Vision and Language

LXMLS 2024

Desmond Elliott

Department of Computer Science

University of Copenhagen

Slides: https://elliottd.github.io/vlprimer/

Working Definition

Multimodal models jointly processes information from two or more input modalities, e.g. images and text, speech and video, etc.

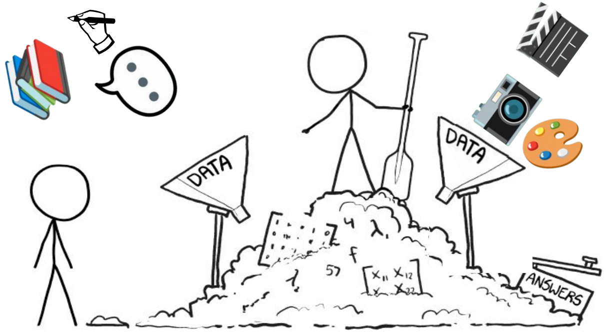
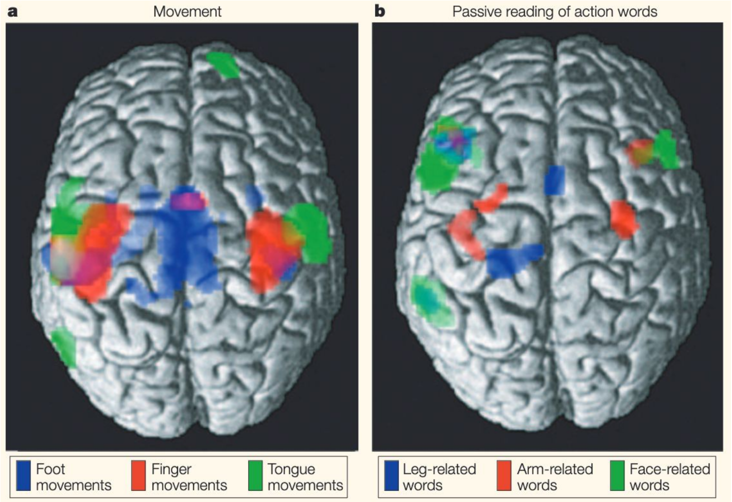


Image adapted from https://xkcd.com/1838/ (CC BY-NC 2.5)

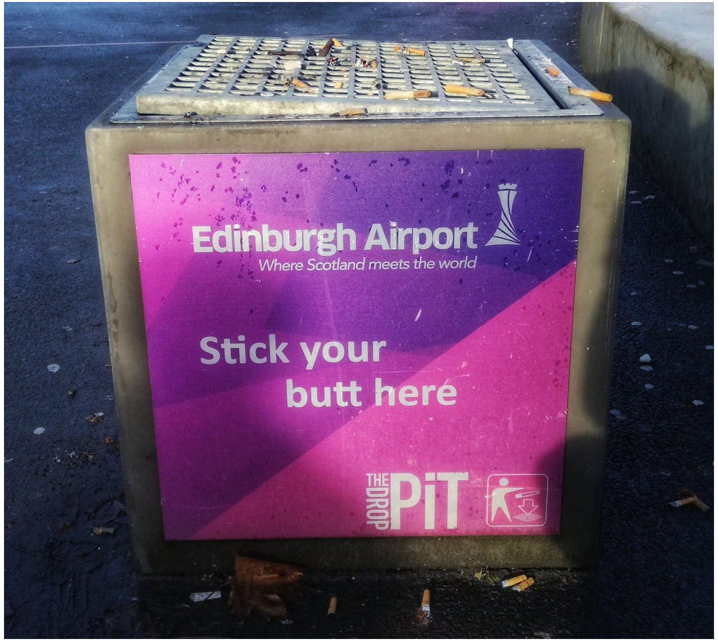
Why Multimodality?

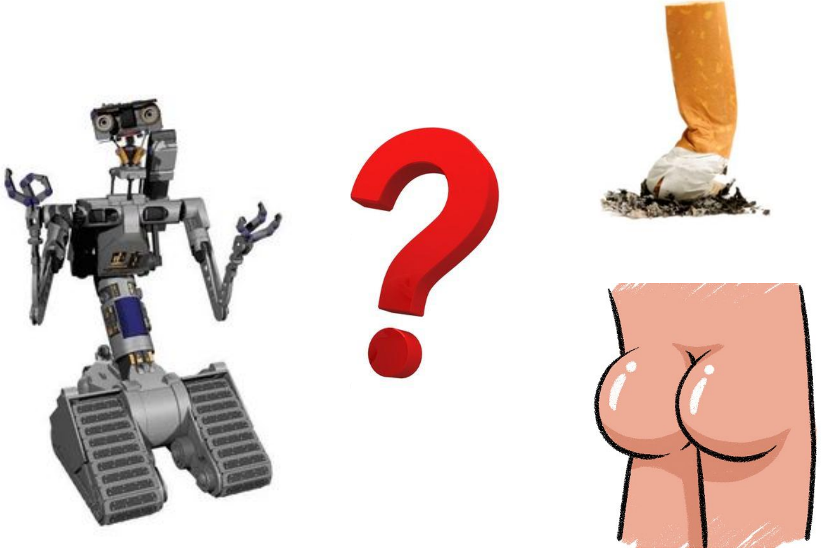
- Humans ground conceptual knowledge in modality processing systems in the brain
- Evidence that grounding activates similar brain regions for different input modalities



Barsalou et al. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in cognitive sciences, 7(2):84–91.
Pulvermüller. (2005). Brain mechanisms linking language and action. Nature reviews neuroscience, 6(7), 576-582.

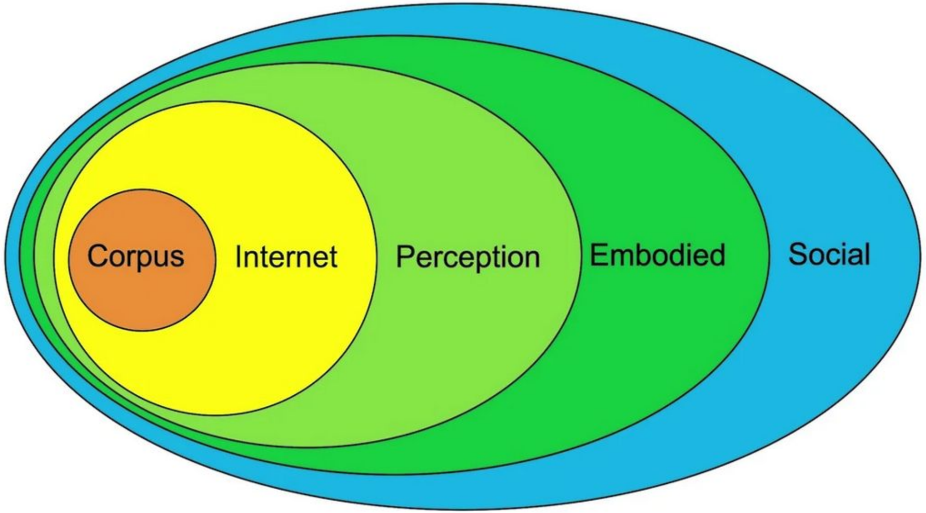
Multimodality reduces ambiguity





You Cannot Learn Language From

- The radio without grounding *(lack perception)*
- The television without actions *(lack embodiment)*
- *W*ithout interacting with others *(lack social)*



Bisk et al. EMNLP 2020.. Experience Grounds Language.

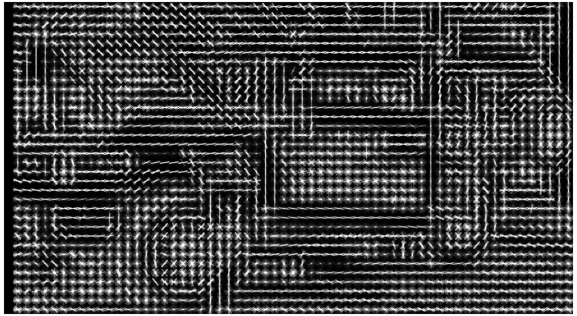
(At Least) Five Major Areas

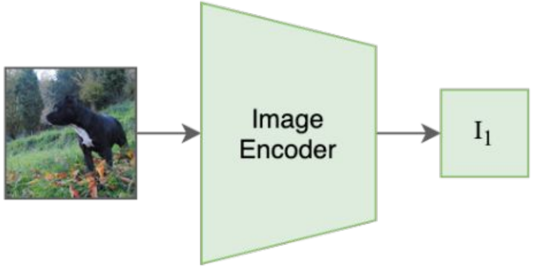
- **Representation**: how to convert raw inputs into a usable format
- **Translation**: transform from one modality to another
- **Alignment**: predict relationships between elements across modalities
- **Fusion**: join features from modalities to support prediction
- **Co-learning**: transferring knowledge from one modality to another

Baltrušaitis, Ahuja, and Morency, (2018). Multimodal machine learning: A survey and taxonomy. IEEE PAMI, 41(2), 423-443.

Representation

- Great deal of work over the last decade, from HOG features in the early 2000s to CLIP features in the 2020s.

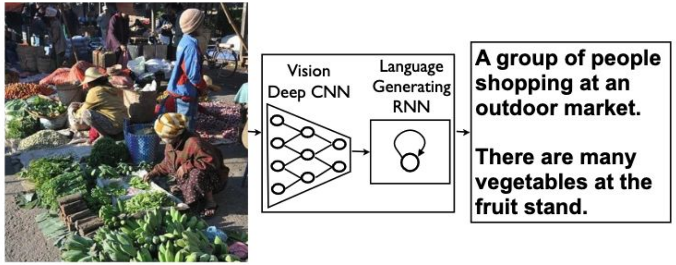


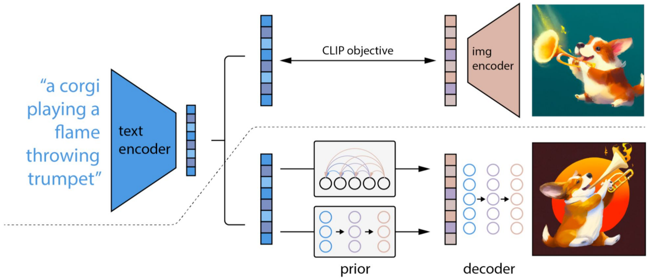


Dalal & Triggs. CVPR 2005. Histograms of oriented gradients for human detection.
Radford et al. ICML 2021. Learning transferable visual models from natural language supervision.

Translation

- Explosion of end-to-end neural network models since the mid 2010s

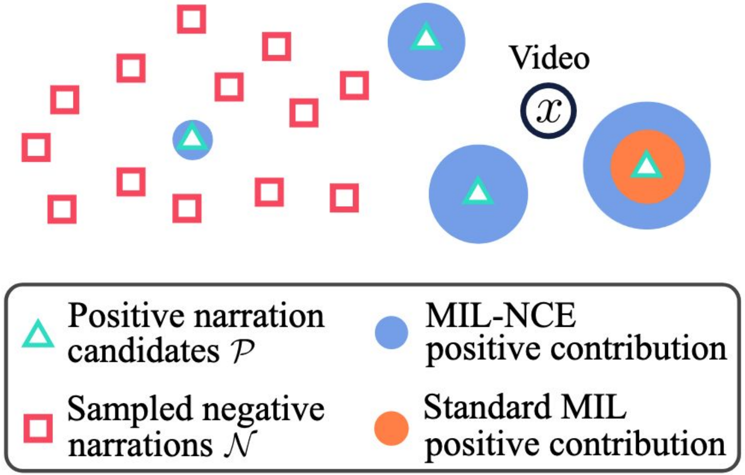

Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.



Vinyals et al. (2015). Show and tell: A neural image caption generator. CVPR.
Ramesh et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.

Alignment

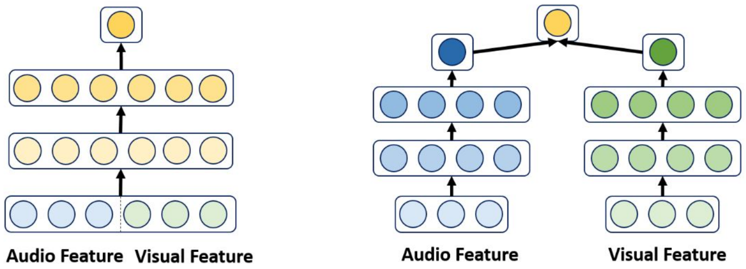
- Important for self-supervised learning and also for phrase grounding

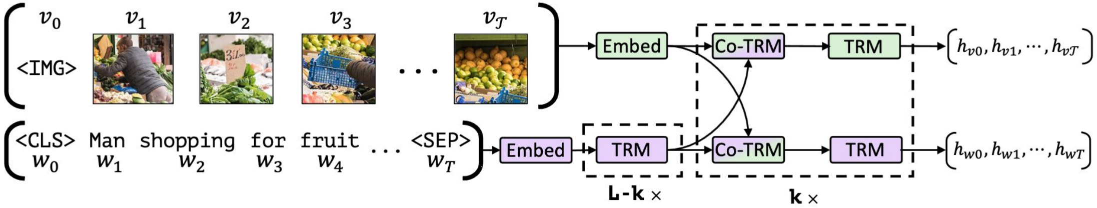

Video
x
Positive narration candidates $\mathcal{P}$
MIL-NCE positive contribution
Sampled negative narrations $\mathcal{N}$
Standard MIL positive contribution

Miech et al. (2020). End-to-end learning of visual representations from uncurated instructional videos. CVPR.
Kamath et al. (2021). MDETR-modulated detection for end-to-end multi-modal understanding. ICCV.

Fusion

- Early work studied the differences between early and late fusion.
- Multi-head self-attention now provides model-based fusion.

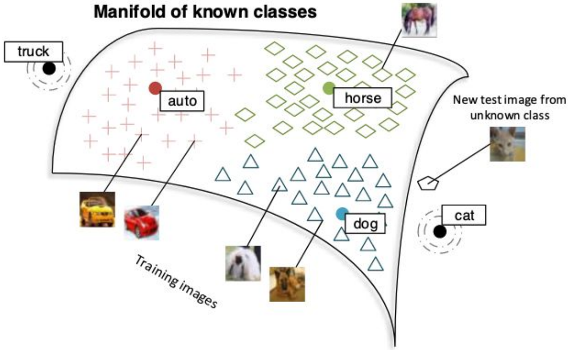


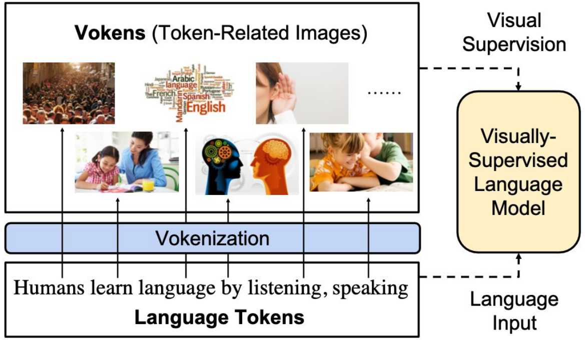


Chen and Jin (2016). Multi-modal conditional attention fusion for dimensional emotion prediction. MM.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS.*

Co-learning

- Zero-shot transfer across modalities, or using visual grounding to improve language models on text-only tasks.




Vokens (Token-Related Images)
Visual Supervision
Visually-Supervised Language Model
Vokenization
Humans learn language by listening, speaking
Language Tokens
Language Input
……

Socher et al. (2013). Zero-shot learning through cross-modal transfer. NeurIPS.
Tan & Bansal. (2020). Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP

Roadmap

Part 1

1. Datasets and Tasks for Multimodal Learning

   📚 Visually Grounded Reasoning across Languages and Cultures

2. Data Representation
3. Modelling Techniques

   📚 Retrieval-Augmentation for Image Captioning, Sequential Multimodal Compositional Generalization

Part 2

4. Understanding Multimodal Models

5. Future Directions

   📚 Language Modelling with Pixels

1. Datasets and Tasks for Multimodal Learning

Two Types of Dataset

- **General-purpose**: visual data with descriptive annotations
    - Conceptual Captions
    - LAION-2/5B
    - Speech-COCO

Blue Beach Umbrellas, Point Of Rocks, Crescent Beach, Siesta Key - Spiral Notebook

- **Task-specific**: visual data with e.g. classification labels
    - Image / Video Captioning
    - Visual Question Answering
    - Visually Grounded Reasoning

What color is the cat's leash?

Many Types of Tasks

- Sequence generation $P(x|v)$
  - Image captioning, video captioning
    visual storytelling, image generation
- Classification $P(y|x, v)$
  - VQA, Visually-grounded Reasoning
- Ranking and Alignment $\mathbf{Distance}(x, v)$
  - Image↔Text Retrieval
    Referring Expression Localization

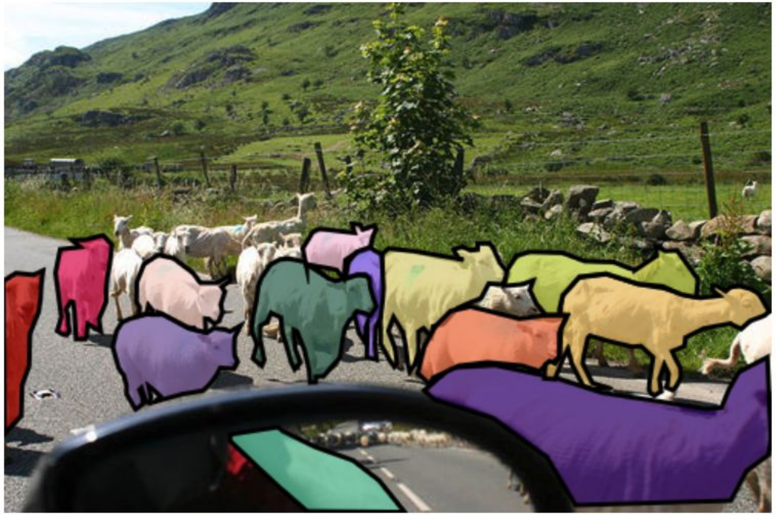
$P(x|v)$

$\mathbf{Distance}(x, v)$

COCO

- Used both a **general-purpose** and **task-specific** dataset
- Images covering 80 *co*mmon objects in *co*ntext with multiple human-authored captions.
- Object segmentation data too!

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.

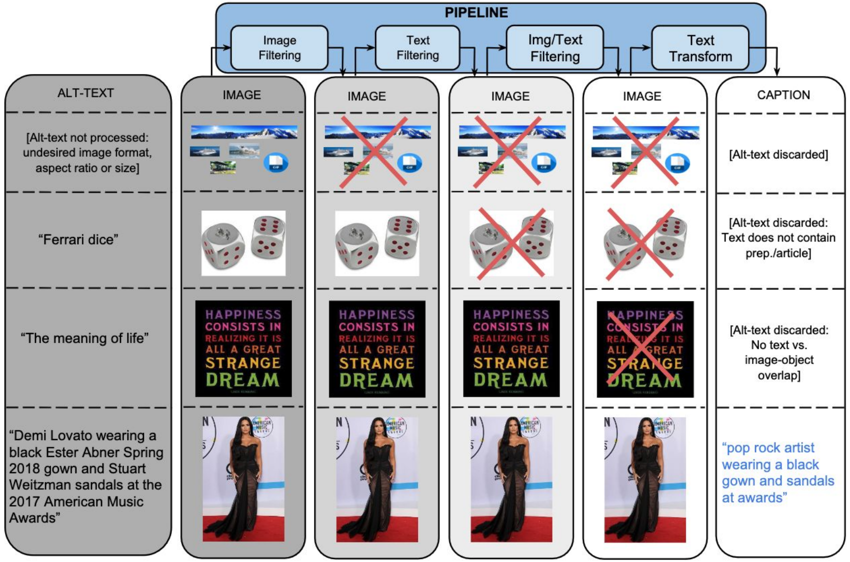


Chen et al. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv.*

$P(x|v)$

Conceptual Captions

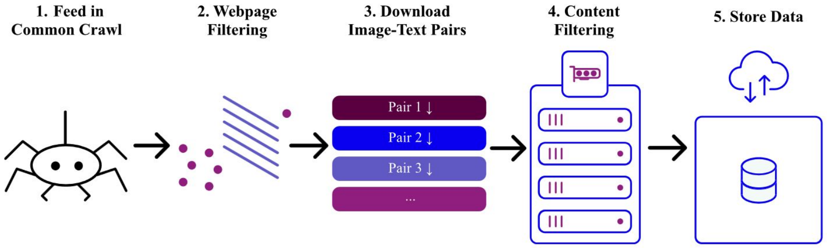
- Used for pretraining
- 3/12M images released with *normalized* English captions.
- Normalization is not public.
- Due to *linkrot*, much less data is currently available.

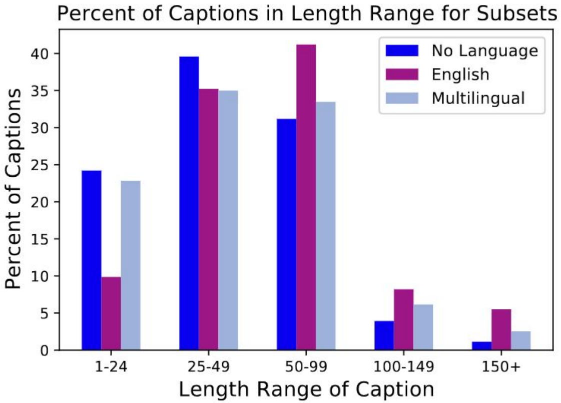

PIPELINE
Image Filtering
Text Filtering
Img/Text Filtering
Text Transform
ALT-TEXT
IMAGE
IMAGE
IMAGE
IMAGE
CAPTION
[Alt-text not processed: undesired image format, aspect ratio or size]
[Alt-text discarded]
"Ferrari dice"
[Alt-text discarded: Text does not contain prep./article]
"The meaning of life"
[Alt-text discarded: No text vs. image-object overlap]
"Demi Lovato wearing a black Ester Abner Spring 2018 gown and Stuart Weitzman sandals at the 2017 American Music Awards"
"pop rock artist wearing a black gown and sandals at awards"
HAPPINESS CONSISTS IN REALIZING IT IS ALL A GREAT STRANGE DREAM

Sharma et al. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. ACL.
Changpinyo et al. (2021). Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. CVPR.
Download your own: https://github.com/igorbrigadir/DownloadConceptualCaptions

LAION

- Used for pretraining
- Image and multilingual *raw* captions harvested from within Common Crawl
- Data behind Stable Diffusion and OpenCLIP
- 5B variant removed due to illegal material



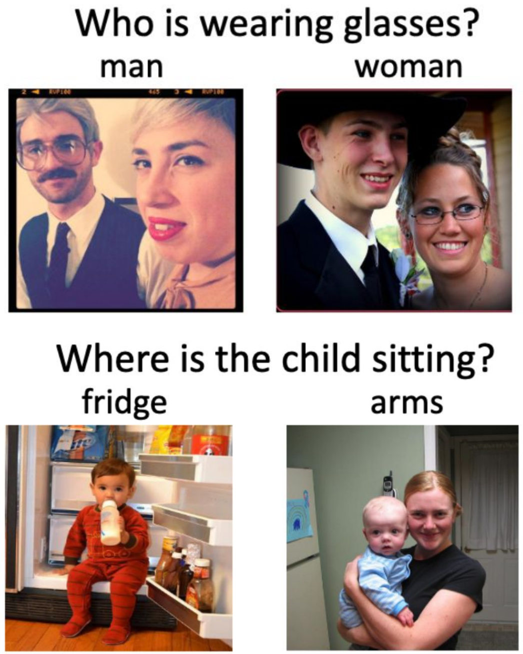


Schuhmann et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. NeurIPS.
Thiel. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models

$P(y|x, v)$

VQAv2

- Answer questions about images
- Task with multimodal inputs:
  - Image
  - Question
- Commonly tackled as classification but increasing efforts as NLG
- 1.1M image–question pairs with balanced distribution of answers

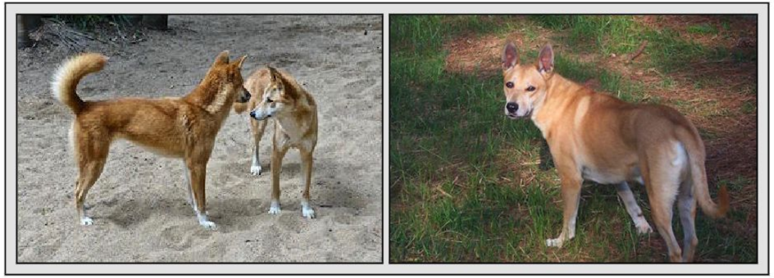


Goyal et al. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. CVPR.

$P(y|x, v)$

NLVR2

- Binary classification task that requires jointly reasoning over a pair of images and a sentence.
- Human-created hard negatives.
- 107K examples in total.



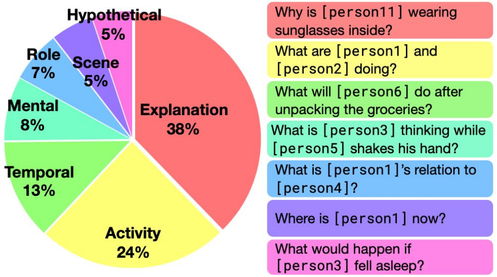
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

Suhr et al. (2019). A Corpus for Reasoning about Natural Language Grounded in Photographs. ACL.

$P(y|x, v)$

Visual Commonsense Reasoning

- 290,000 multiple-choice VQA examples derived from movies.



Why is [person11] wearing sunglasses inside?

What are [person1] and [person2] doing?

What will [person6] do after unpacking the groceries?

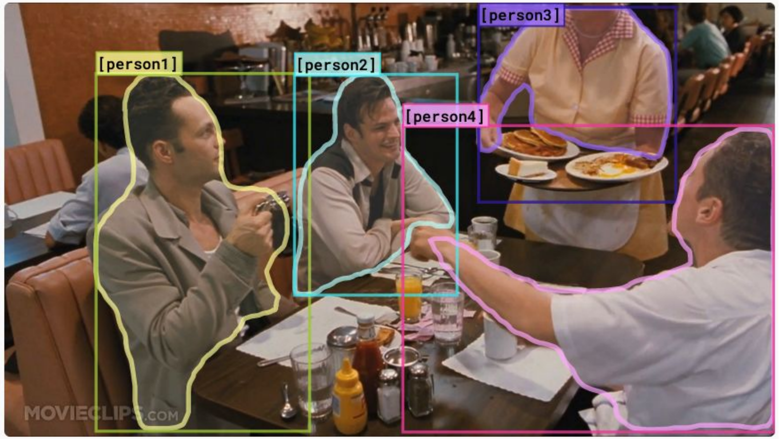
What is [person3] thinking while [person5] shakes his hand?

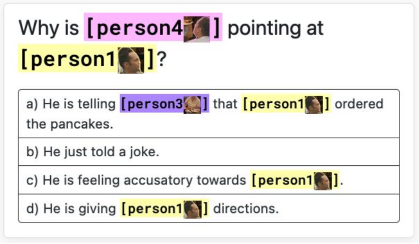
What is [person1]'s relation to [person4]?

Where is [person1] now?

What would happen if [person3] fell asleep?

- In addition to Question Answering, the dataset includes rationale selection too!





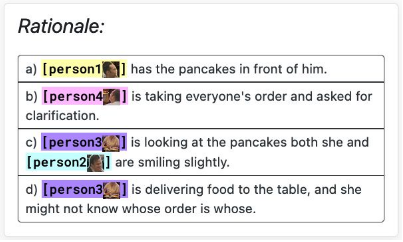
Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.



Rationale:

a) [person1] has the pancakes in front of him.

b) [person4] is taking everyone's order and asked for clarification.

c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.

d) [person3] is delivering food to the table, and she might not know whose order is whose.

Zellers et al. (2019). From Recognition to Cognition: Visual Commonsense Reasoning. CVPR.

$P(y|x, v)$
$\mathbf{Distance}(x, v)$

Multi30K

- Multilingual aligned image–sentence dataset in many languages
  - English, German, French, Czech, Arabic, Japanese, Turkish, Ukranian

A group of people are eating noodles.

Eine Gruppe von Leuten isst Nudeln.

Un groupe de gens mangent des nouilles.

Skupina lidí jedí nudle.

Elliott et al. (2016). Multi30K: Multilingual English-German Image Descriptions. ACL Workshop.

$P(y|x,v)$
$\mathbf{Distance}(x,v)$

BOBSL

- BBC-Oxford British Sign Language Dataset
- Sign spotting and sentence localization tasks
- 1,400 hours of signed shows
    - Factual, entertainment, drama, comedy, children's shows

Albanie et al. (2021). BBC-Oxford British Sign Language Dataset. arXiv.

Many Many More

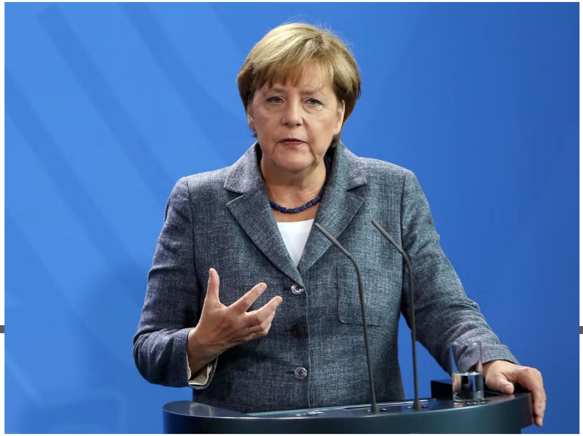
- Visual Storytelling, e.g. VIST
- Grounded Referring Expression, e.g. Flickr30K Entities, Visual Genome
- Visual Entailment, e.g. SNLI-VE
- Vision & Language Navigation, e.g. RxR
- Visual Common Sense Reasoning: VCR
- Text-to-Image Generation, e.g. DALLEval
- Abstract reasoning, e.g. KiloGram, CRAFT
- Sign Language Processing, e.g. How2Sign

- *and more and more and more and more*

Binding: Degree of Multimodality

- The content expressed in textual data depends on the purpose

Social media platforms often form 'echo chambers' that encourage users to only read content that confirms beliefs they already hold (Getty)



A woman in a grey suit is giving a speech

Weak ← → Strong

(Crawled)

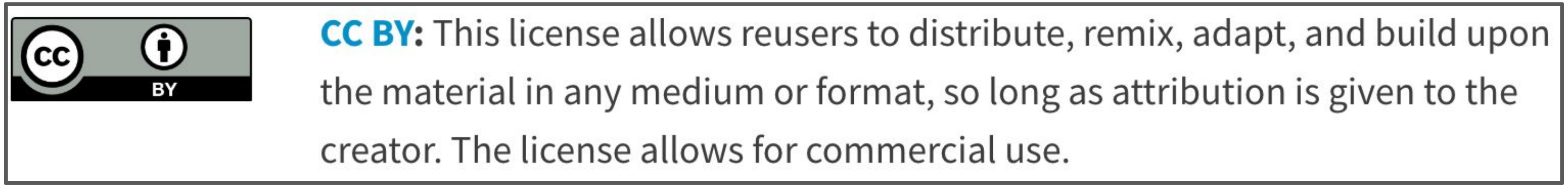
(Crowdsourcing)

Rewriting crawled text improves performance on a variety of downstream multimodal tasks

Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022.
Nguyen et al. Improving multimodal datasets with image captioning. NeurIPS 2023.

Ethical Issues

- Multimodal datasets are usually data scraped from the web with *unknown degrees of conformance*, or information about, licensing.



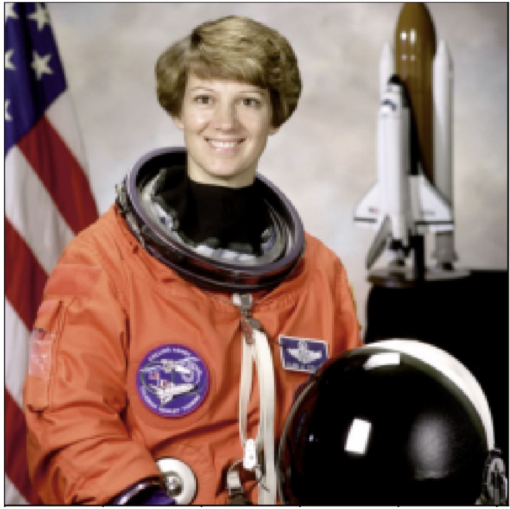
CC BY: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

- As of 2022, there are an estimated 2.5B CC-licensed objects online.

State of the Commons 2022. Creative Commons Foundation.

The Problem with Scale

- Scale lets you build systems that perpetuate harmful stereotypes



(Eileen Collins, American astronaut)

`cos(v,x)`

0.276 ← "This is a portrait of an astronaut with the American flag"

0.308 ← "This is a photograph of a smiling housewife in an orange jumpsuit with the American flag"

Birhane et al. (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv.

Q: How can we collect multimodal data that better reflects the diversity of the world?

Visually Grounded Reasoning across Languages and Cultures

EMNLP 2021

F. Liu*

E. Bugliarello*

E.M. Ponti

S. Reddy

N. Collier

D. Elliott

Motivation

Languages

- Mostly in English
- Or some Indo-European Languages

ENG: An unusual looking vehicle …

NLD: Een mobiel draaiorgel …

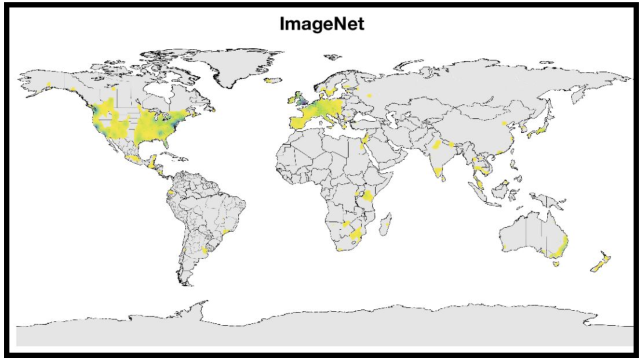
Example from van Miltenburg+ 2017

Image sources

- Mostly from ImageNet or COCO
- Reflecting North American and European cultures

Implications for V&L models

- Narrow linguistic/cultural domain
- No way to assess their real-world comprehension



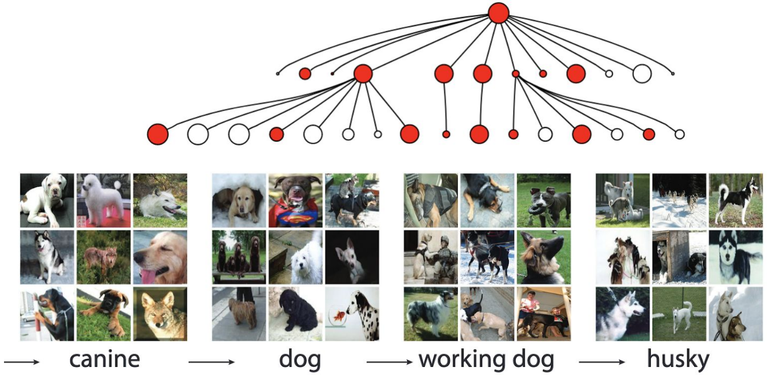
Density map of geographical distribution of images in ImageNet (DeVries+, 2019)

F. Liu*, E. Bugliarello*, et al. Visually Grounded Reasoning across Languages and Cultures. EMNLP 2021.

Typical Vision and Language DATA

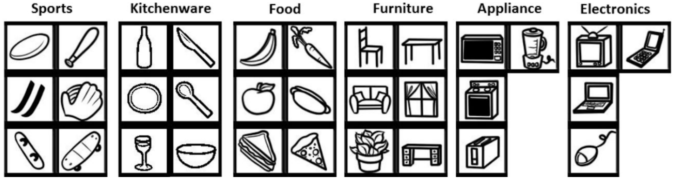
ImageNet (Deng et al. 2009)

- Train visual encoders
- Millions of labelled images
- Derived from the WordNet hierarchy



Common Objects in Context (Lin et al. 2014)

- Train and evaluate multimodal models
- 330K labelled images
  - 80 types of commonly occurring objects



Concrete Concepts in Cultural Context

- Some concepts are most immediately understood within a cultural background

Culture: The way of life of a collective of people that distinguishes them from other people (Mora, 2013; Shweder et al. 2007).

Pilota / Jai-alai

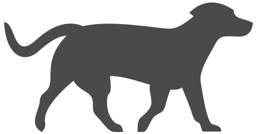
Sanxian / Shamisen

Clavie

Concepts and Hierarchies

Category: objects with similar properties (Aristotle 40 BCE, …)

Concept: mental representation of a category (Rosch 1973)



“Dog” concept

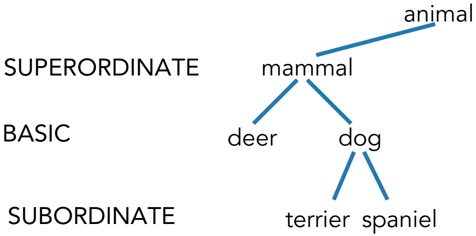
“Dog” category

Categories form a *hierarchy*

- Basic-level categories (Rosch 1976)

Somewhat universal

- Different cultures (Berlin 2014)
- Familiarity of individuals (Wisniewski and Murphy, 1989)



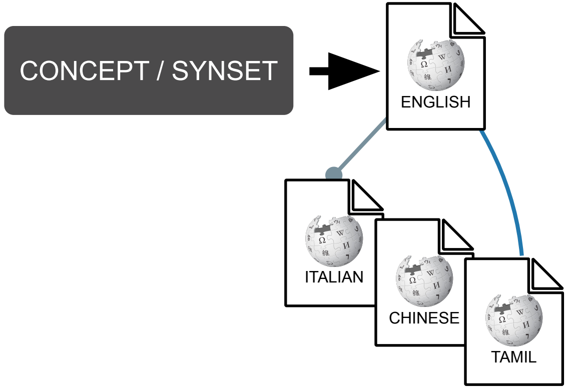
SUPERORDINATE

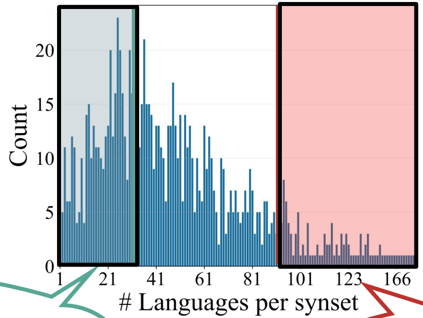
BASIC

SUBORDINATE

Are ImageNet Concepts Cross-Lingual?

- ImageNet, COCO and Visual Genome use English WordNet concepts
- Question: estimate cross-linguality using Wikipedia as a proxy





Most concepts in <30 languages

Only a few "universal" concepts

MaRVL Multicultural Reasoning over Vision and Language

Representative of annotators' cultures

5 typologically diverse languages

Independent, culture-specific annotations

MaRVL-id Bola basket

MaRVL-sw Mpira wa kikapu

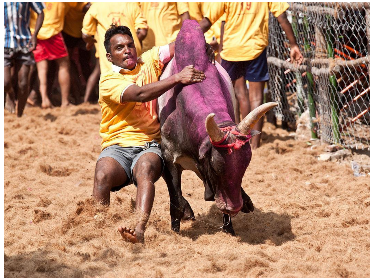
MaRVL-tr Basketbol

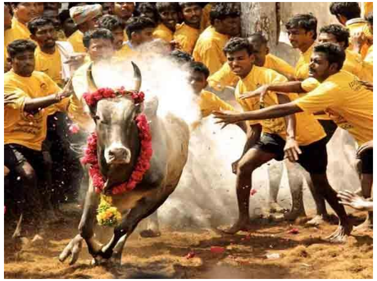
MaRVL-zh 篮球

MaRVL-ta கூடைப்பந்தாட்டம்

Visual Reasoning Task

- **Datapoint**: two images (v_1, v_2) paired with a sentence x
- **Task**: Predict whether x is a true description of the pair of images v_1 v_2





இரு படங்களில் ஒன்றில்
இரண்டிற்கும் மேற்பட்ட
மஞ்சள் சட்டை அணிந்த
வீரர்கள் காளையை அடக்கும்
பணியில் ஈடுப்பட்டிருப்பதை
காணமுடி.

True

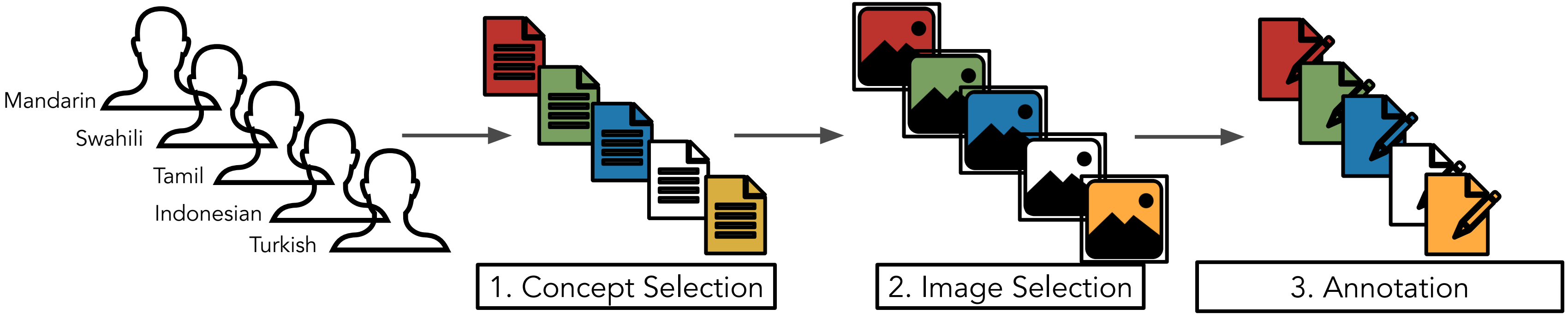
x

y

Suhr et al. A Corpus for Reasoning about Natural Language Grounded in Photographs. ACL 2019.

Collecting MaRVL data

Native speaker-driven protocol



Mandarin

Swahili

Tamil

Indonesian

Turkish

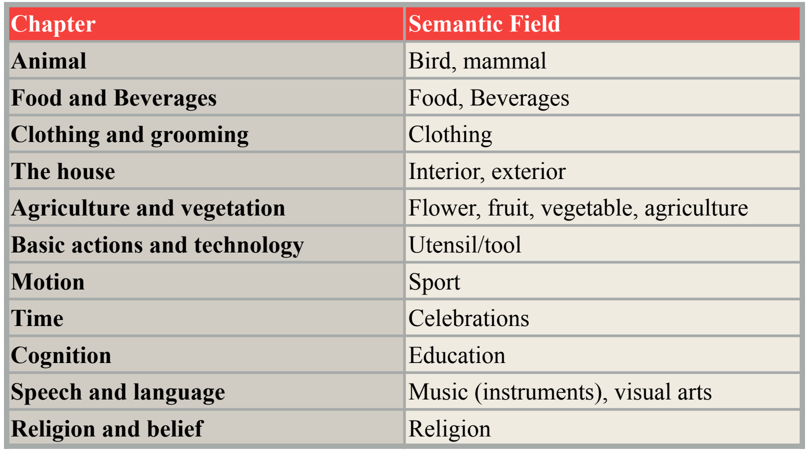
1. Concept Selection

2. Image Selection

3. Annotation

MaRVL is created from Universal Concepts

- Taken from the *Intercontinental Dictionary Series* (Key & Comrie, 2015)
  - 18/22 chapters with concrete objects & events



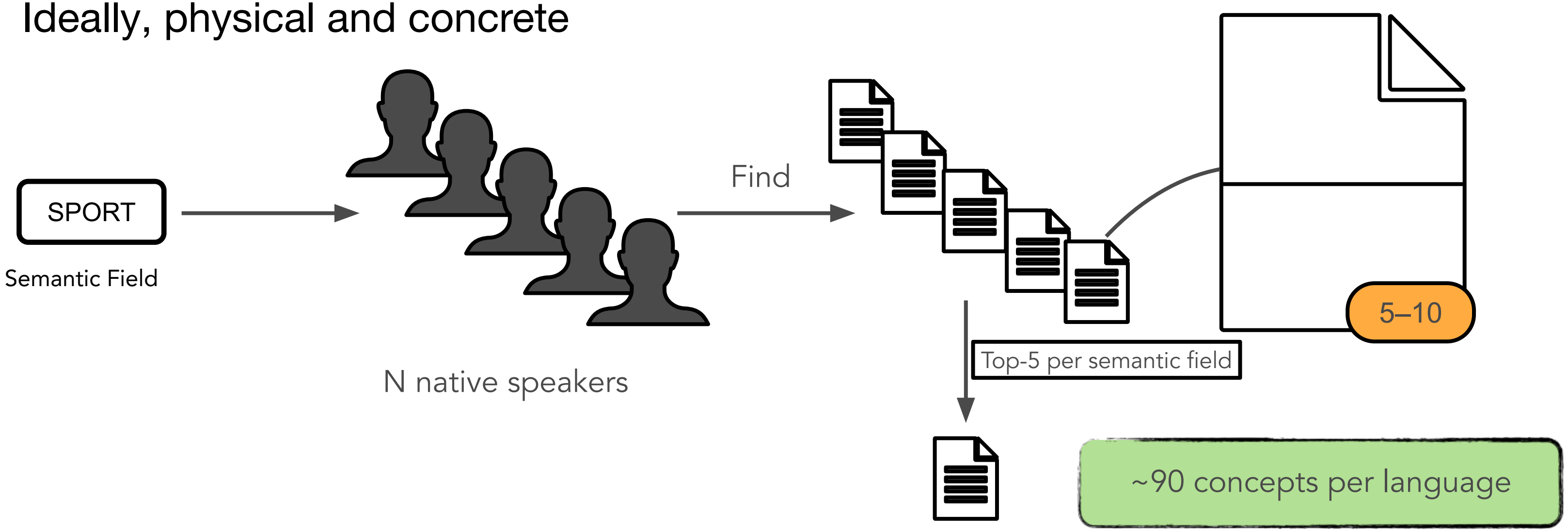
Chapter	Semantic Field
Animal	Bird, mammal
Food and Beverages	Food, Beverages
Clothing and grooming	Clothing
The house	Interior, exterior
Agriculture and vegetation	Flower, fruit, vegetable, agriculture
Basic actions and technology	Utensil/tool
Motion	Sport
Time	Celebrations
Cognition	Education
Speech and language	Music (instruments), visual arts
Religion and belief	Religion

OUR BIAS

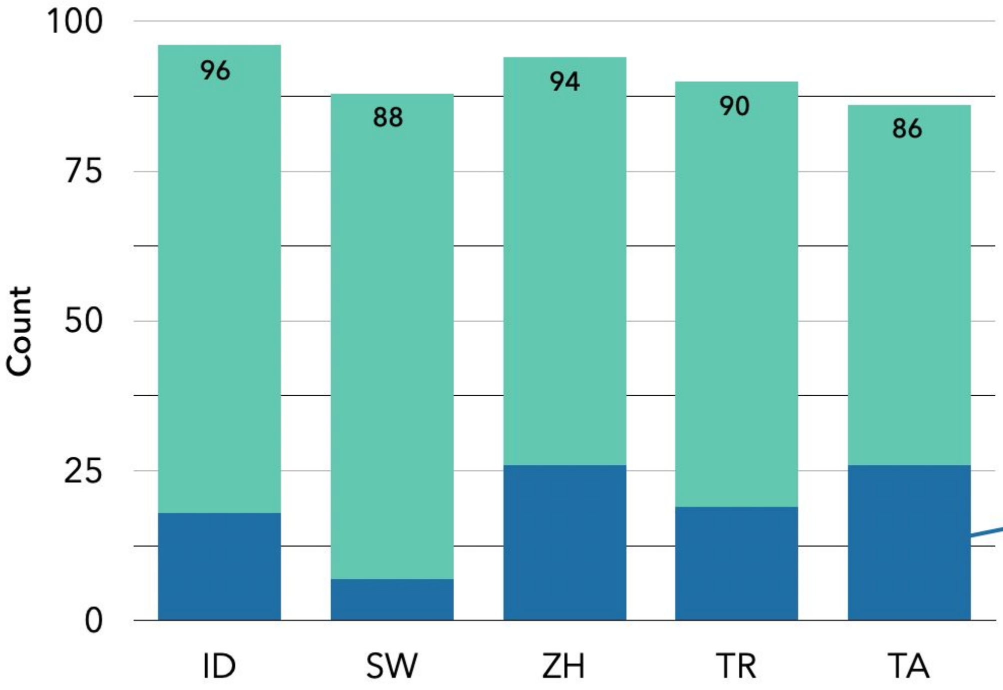
Step 1. Language-Specific Concepts

Defined by native speakers

- Commonly seen or representative in their culture
- Ideally, physical and concrete



Overview of Resulting Concepts



NOT IN ENGLISH WORDNET!

Step 2. Image Collection

Collected by native speakers

- Representative of the language population
- NLVR2 (Suhr et al. ACL 2019) requirements
  1. Contains more than one instance of a concept
  2. Shows an instance of the concept interacting with other objects
  3. Shows an instance of the concept performing an activity
  4. Displays a set of diverse objects or features

MaRVL-ta மோர் (Buttermilk)

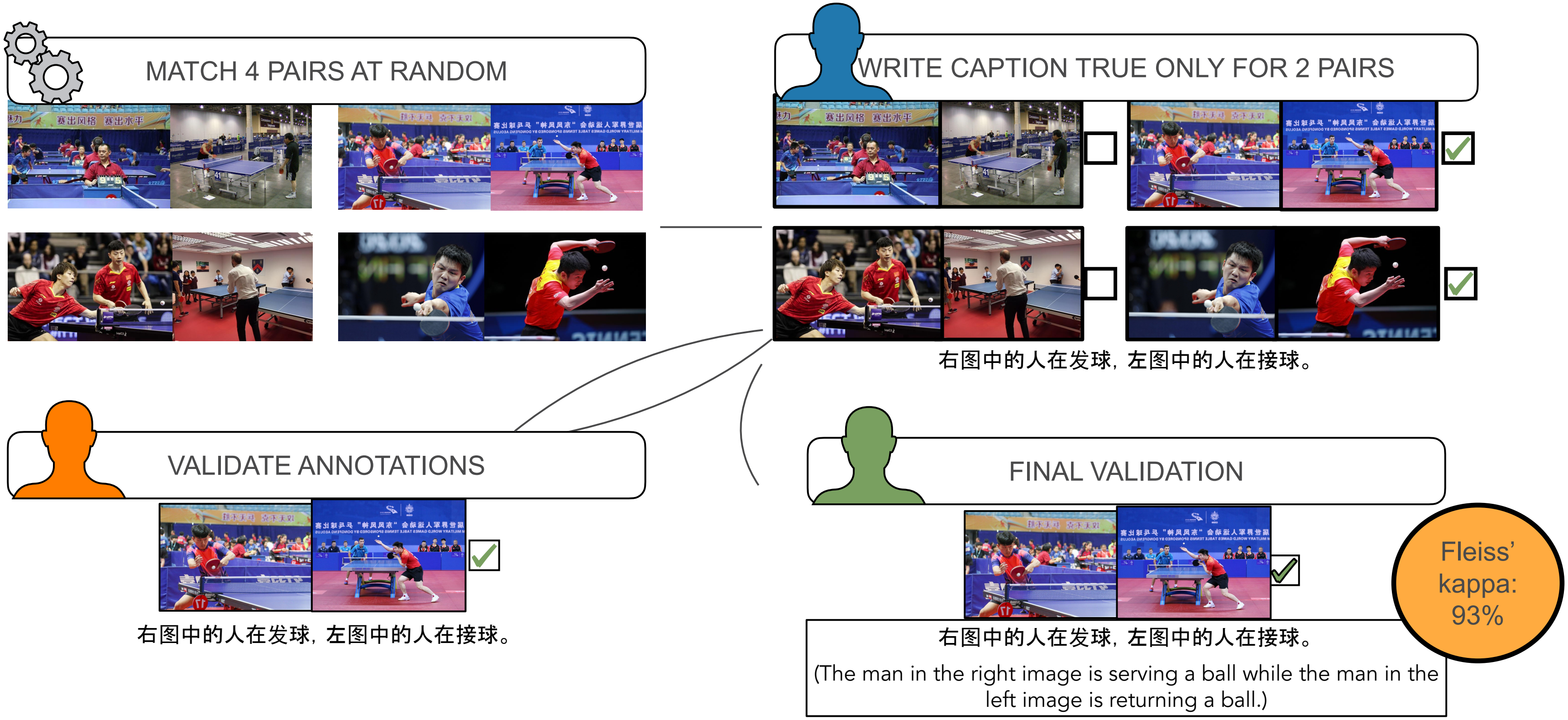
MaRVL-zh 花椰菜 (Cauliflower)

MaRVL-tr Rakı (Raki)

MaRVL-sw Jembe (Shovel)

Step 3. Language Annotation

Written by native speakers



MATCH 4 PAIRS AT RANDOM

WRITE CAPTION TRUE ONLY FOR 2 PAIRS

☐ ☑

☐ ☑

右图中的人在发球，左图中的人在接球。

VALIDATE ANNOTATIONS

☑

右图中的人在发球，左图中的人在接球。

FINAL VALIDATION

☑

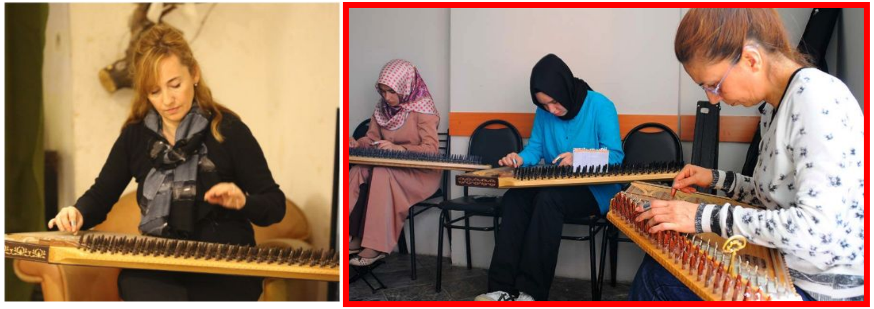
右图中的人在发球，左图中的人在接球。

(The man in the right image is serving a ball while the man in the left image is returning a ball.)

Fleiss' kappa: 93%

Dataset Examples

MaRVL-tr Kanun (çalgı)

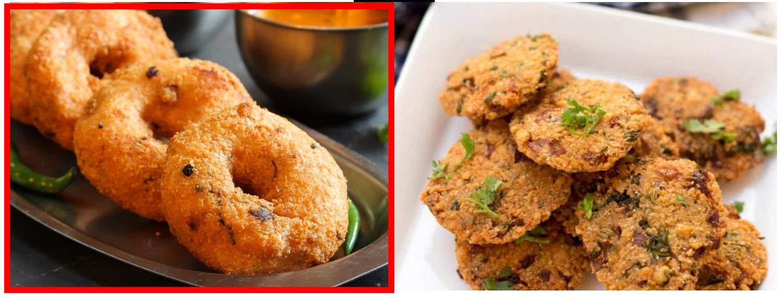


Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: True

MaRVL-ta மை (Vada)

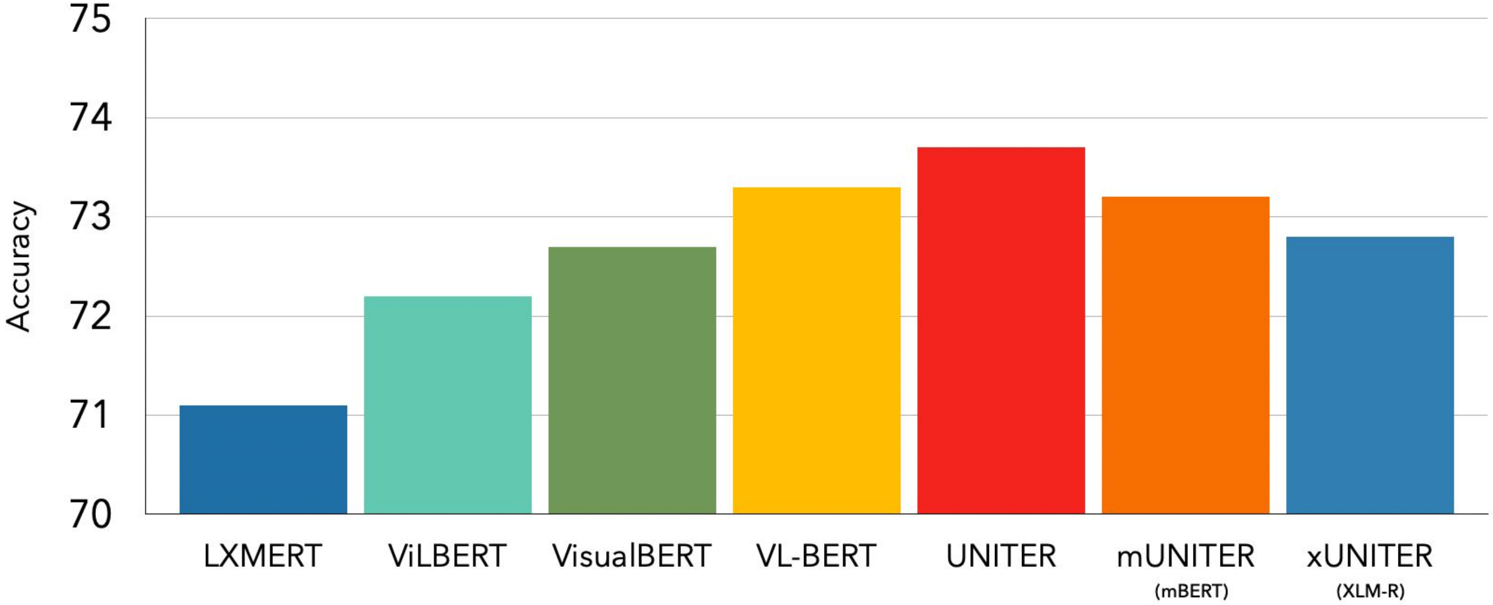


இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்

(Both images contain a lot of masala vadas)

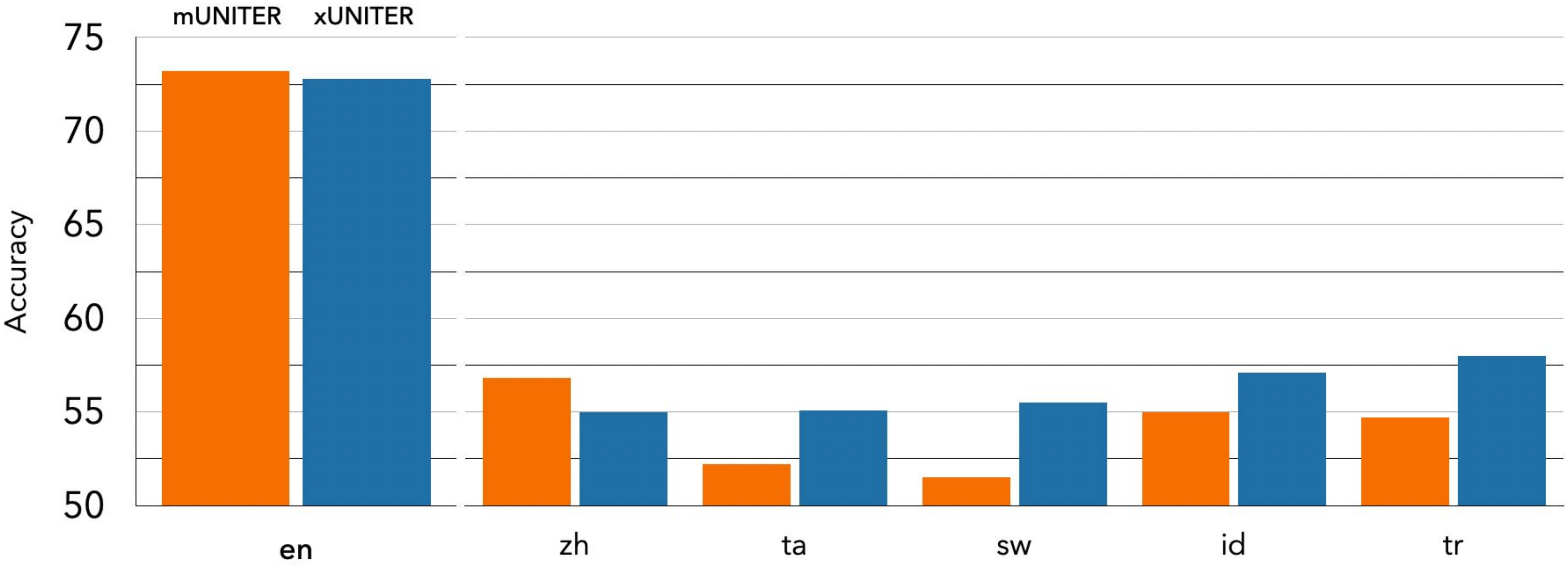
Label: False

English NLVR2 Results (Sanity check)



m/xUNITER perform similarly to English-only models

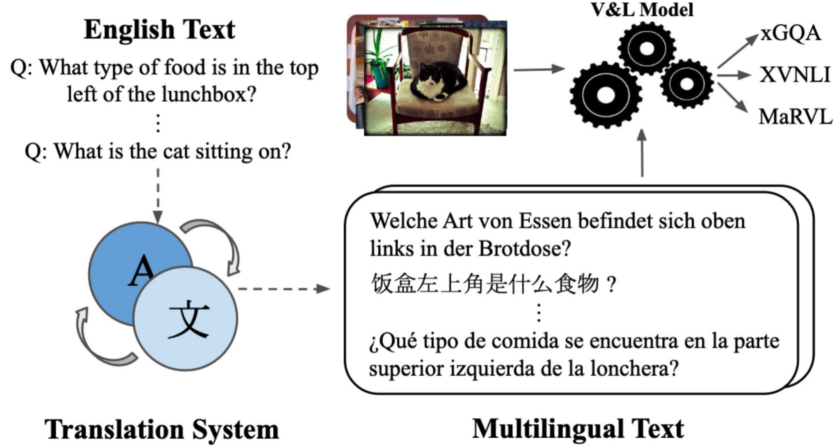
MaRVL Zero-shot Results



Zero-shot transfer: substantial drop in performance

Pretraining with Translated Text

- Are the low zero-shot results caused by poor cross-lingual multimodal binding?



Q: What type of food is in the top left of the lunchbox?

⋮

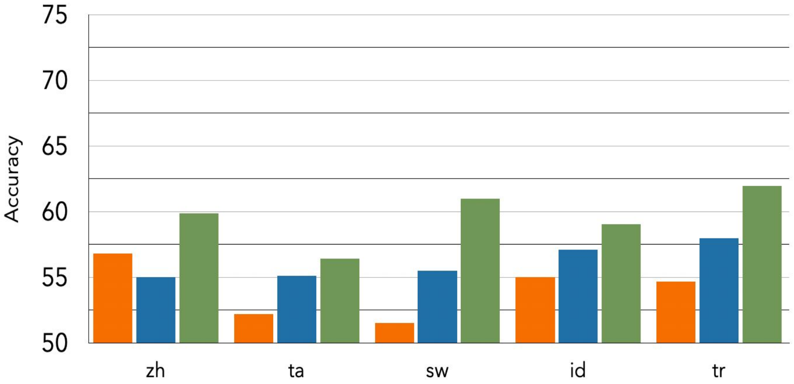
Q: What is the cat sitting on?

Welche Art von Essen befindet sich oben links in der Brotdose?

饭盒左上角是什么食物？

⋮

¿Qué tipo de comida se encuentra en la parte superior izquierda de la lonchera?



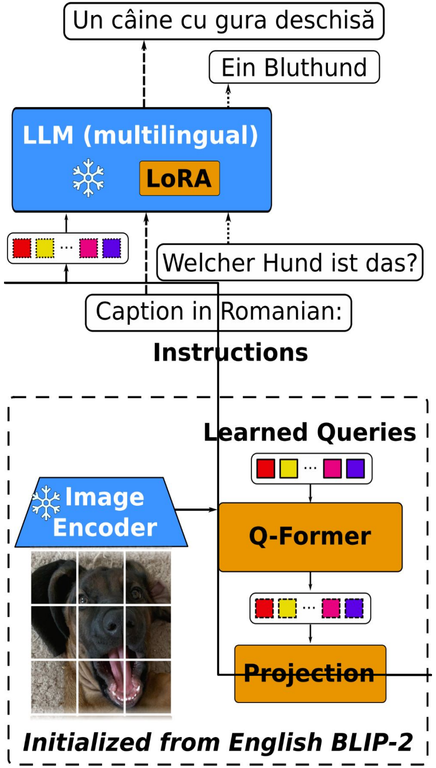
Cross-modal multilingual multimodal pretraining helps!

Qui et al. Multilingual Multimodal Learning with Machine Translated Text. Findings of EMNLP 2022.

Supervised State of the Art: mBLIP

- Initialize from
  - BLIP-2
  - MT0-XL

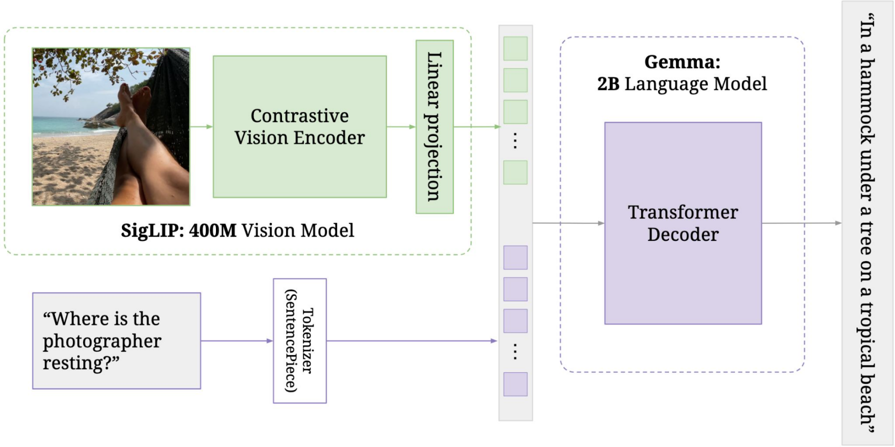
- Use NLLB to pretrain on 96 languages
  - MSCOCO
  - CapFilt
  - VQAV2 & A-OKVQA
  - ImageNet as multilingua VQA



Geigle et al. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. 2023.

Zero-shot State of the Art: PaliGemma

- Initialize from
  - Gemma 2B
  - SigLIP-So400m/14
- Pretrain on
  - Web Language Image (???)
  - CC3M-35L
  - VQ^{2}A/VQG-CC3M-35L
  - OpenImages
  - Wikipedia Image Text

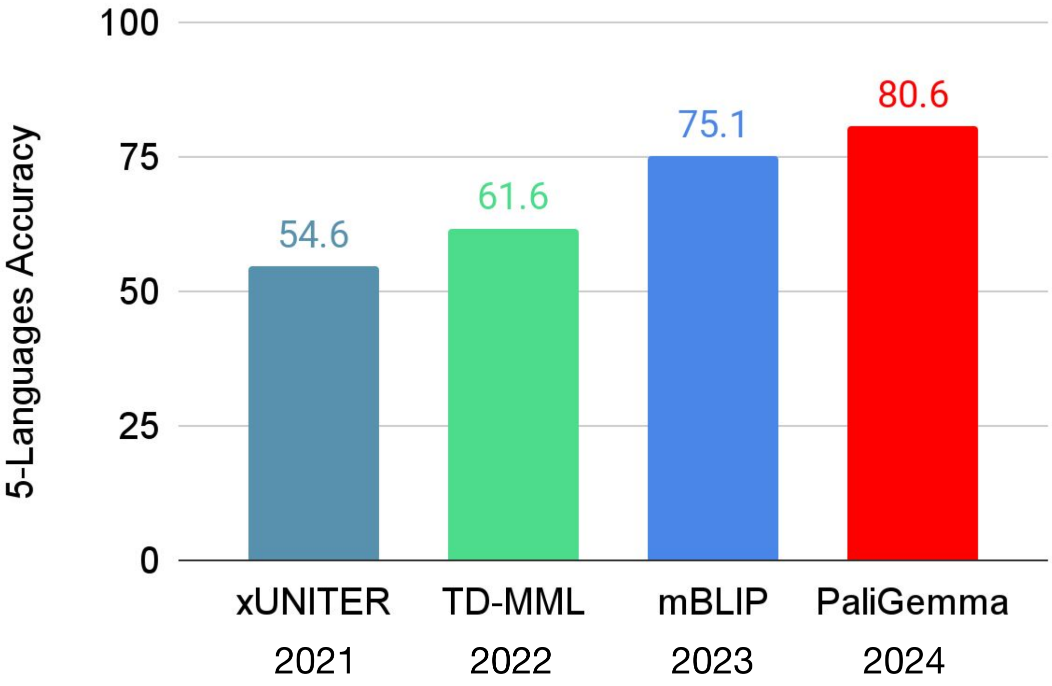


Contrastive Vision Encoder
Linear projection
SigLIP: 400M Vision Model
Gemma:
2B Language Model
Transformer Decoder
"Where is the photographer resting?"
Tokenizer (SentencePiece)
"In a hammock under a tree on a tropical beach"

Year-on-Year Improvements

- Clear benefit when using machine translated data

- Better visual encoders and language models can enable effective zero-shot transfer

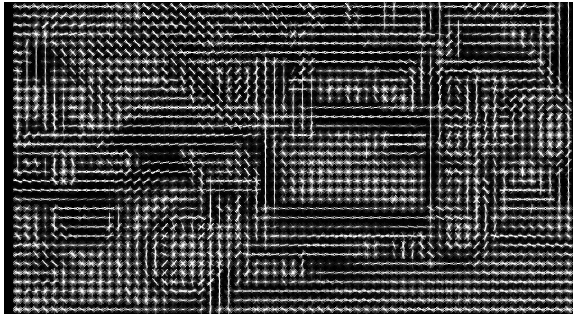


2. Data Representation

Three Levels of Representation

- Perceptual
- Pre-processed features
- Raw input

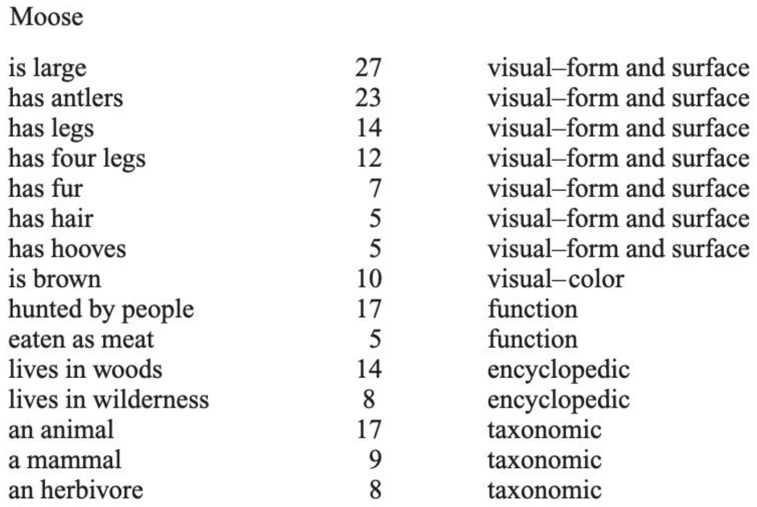
❏ Yellow
❏ Has wheels
❏ Metal
❏ Five-door
❏ Can transport
❏ …



Images: http://www.cs.columbia.edu/~vondrick/ihog/

Perceptual Norms

- Ask people to write down the words that are triggered by textual stimuli.
- Stimuli: 541 noun concepts
- Norms are categorized into the likely knowledge source



Moose

is large	27	visual–form and surface
has antlers	23	visual–form and surface
has legs	14	visual–form and surface
has four legs	12	visual–form and surface
has fur	7	visual–form and surface
has hair	5	visual–form and surface
has hooves	5	visual–form and surface
is brown	10	visual–color
hunted by people	17	function
eaten as meat	5	function
lives in woods	14	encyclopedic
lives in wilderness	8	encyclopedic
an animal	17	taxonomic
a mammal	9	taxonomic
an herbivore	8	taxonomic

McRae et al. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547.

Perceptual Norms: Pros / Cons

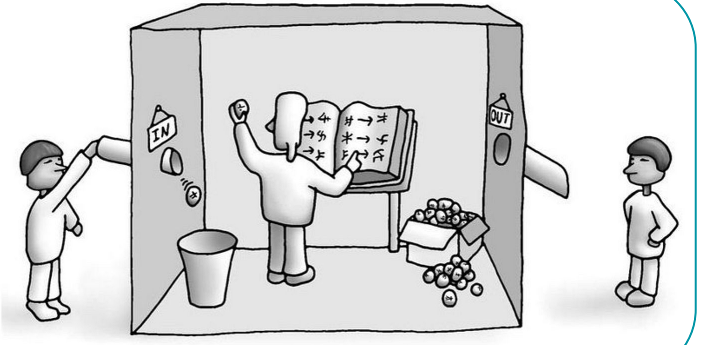
Pros

- Seemingly simple task
- Rich features

Cons

- Can it scale?
- Handling ambiguity

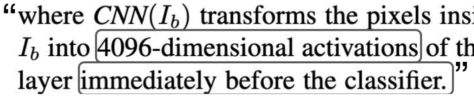
What does it mean to only understand symbols as defined by other symbols?



Searle. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3: 417–57

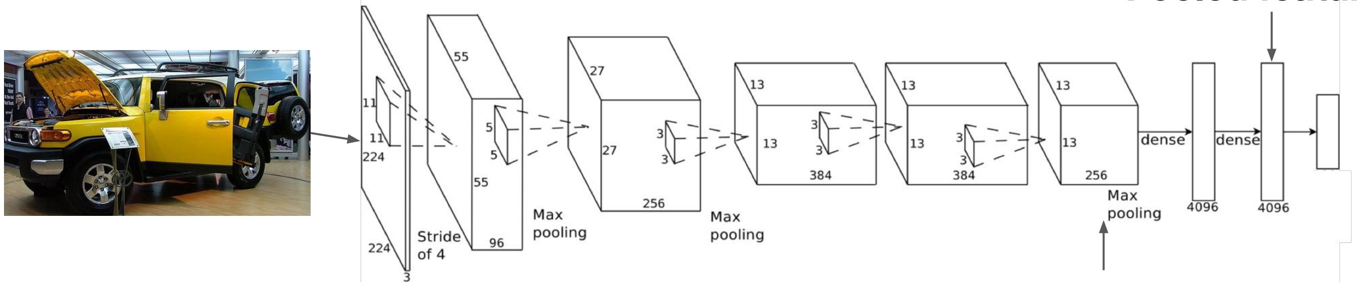
Spatial and Pooled Visual Features

- Earliest work in neural-network era used pooled or spatial preserving features from a pretrained Convolutional Neural Network.

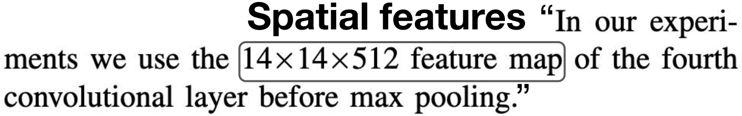


“where $CNN(I_b)$ transforms the pixels inside bounding box I_b into $\boxed{\text{4096-dimensional activations}}$ of the fully connected layer $\boxed{\text{immediately before the classifier.}}$”

Pooled features



Spatial features

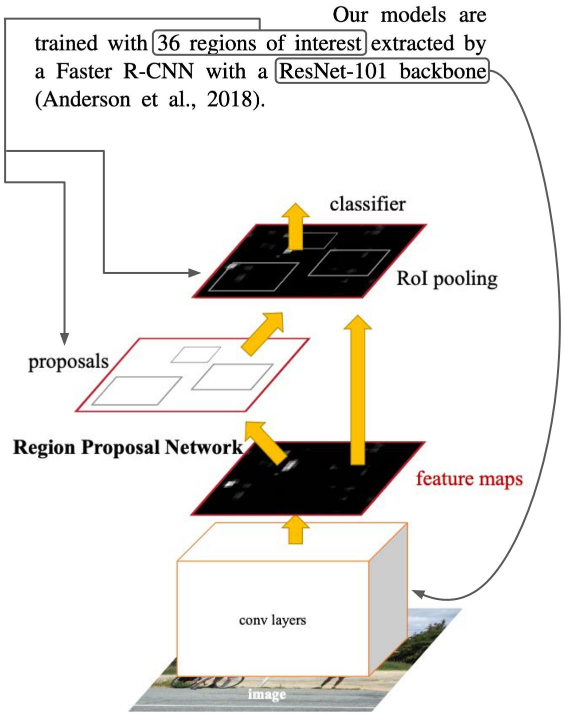


“In our experiments we use the $\boxed{14\times14\times512 \text{ feature map}}$ of the fourth convolutional layer before max pooling.”

Karpathy & Fei-Fei (2015). Deep visual-semantic alignments for generating image descriptions. CVPR.
Xu et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. ICML.

Pre-processed Visual Features

- Faster R-CNN region-based feature vectors
  - Trained on the Visual Genome Dataset
  - The Region Proposal Network suggests the location of *regions of interest*.
  - RoI pooling performs spatial pooling in the final CNN layer to give a 2048D vector.



Our models are trained with 36 regions of interest extracted by a Faster R-CNN with a ResNet-101 backbone (Anderson et al., 2018).

Ren et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS.*
Bugliarello et al. (2021). Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. TACL.

Pre-processed: Pros / Cons

Pros

- Long-established practice
- Usually an offline process: do it once and forget

Cons

- Large datasets require specialized storage
- Not obvious how to randomly augment data
- Specialist knowledge can be opaque to newcomers

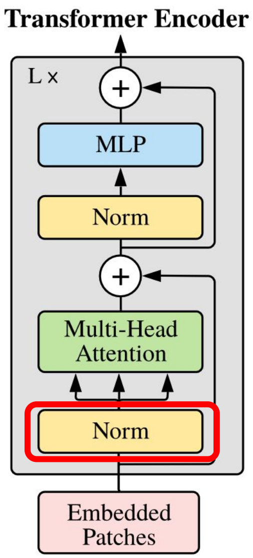
Raw Input

- Directly process data from the raw images or speech signal.
- Images:
  - Vision Transformer (ViT)
  - Swin Transformer
- Speech
  - Spectrogram Transformer
  - AudioMAE

Transformers | Davide Coccomini | 2021

Vision Transformer

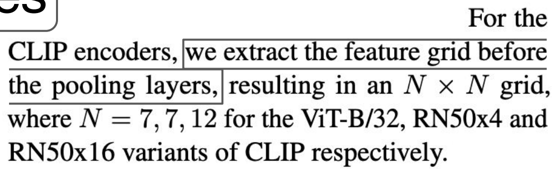
- Good news! You are already almost an expert in how the Vision Transformer works
    - Split image into K patches
    - Embed each patch
    - Add position information
    - Encode using Transformer blocks that include an extra pre-norm layer for stability.



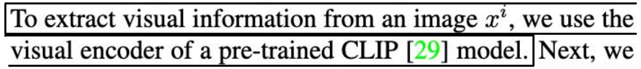
Dosovitskiy et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
Baevski & Auli (2018). Adaptive input representations for neural language modeling. *ICLR.*

Extracting ViT Features

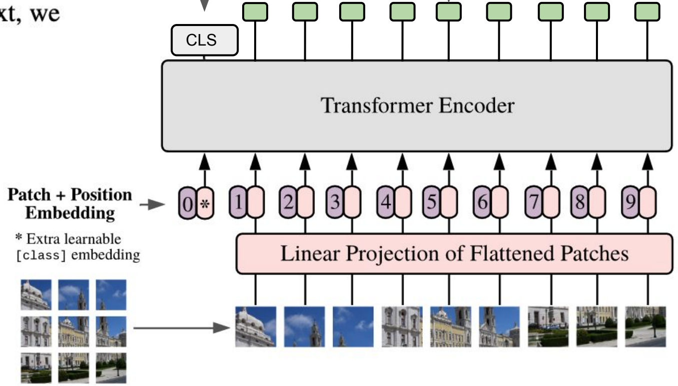
- Extract pooled features or patch-level features



For the CLIP encoders, we extract the feature grid before the pooling layers, resulting in an $N \times N$ grid, where $N = 7, 7, 12$ for the ViT-B/32, RN50x4 and RN50x16 variants of CLIP respectively.



To extract visual information from an image x^i, we use the visual encoder of a pre-trained CLIP [29] model. Next, we



CLS
Transformer Encoder
Patch + Position Embedding
0
*
1
2
3
4
5
6
7
8
9
* Extra learnable [class] embedding
Linear Projection of Flattened Patches

Eichenberg et al. (2021). MAGMA--Multimodal Augmentation of Generative Models through Adapter-based Finetuning. *EMNLP*

Raw input: Pros / Cons

Pros

- Data augmentation is straightforward because you always have the raw input
- Fewer preprocessing steps means fewer creeping errors

Cons

- Smaller batches with an extra model on the GPU
- Potentially many inputs

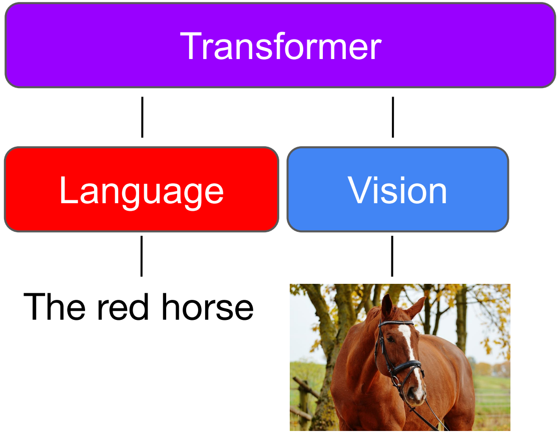
Summary

- Many options for how to represent your multimodal inputs
  - Language-oriented
  - Object / stuff oriented
  - Raw inputs
- **No universally best option** but raw inputs are promising because the visual encoding model can be fully differentiable

3. Modelling

Main Approaches

Dual / cross encoding



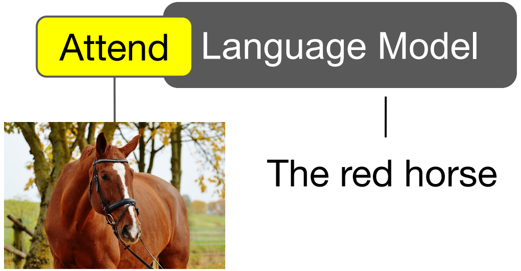
Transformer

Language

Vision

The red horse

Cross-Attention

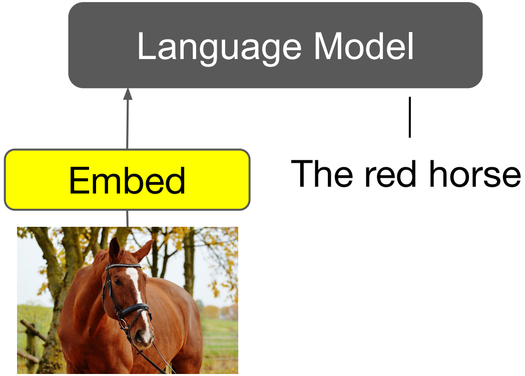


Attend

Language Model

The red horse

Visual Prefix



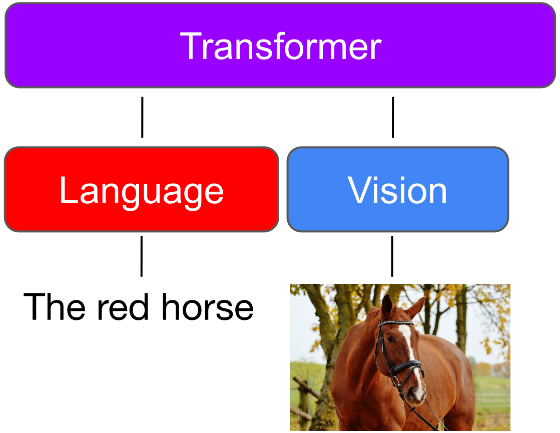
Language Model

Embed

The red horse

Cross-encoding Models

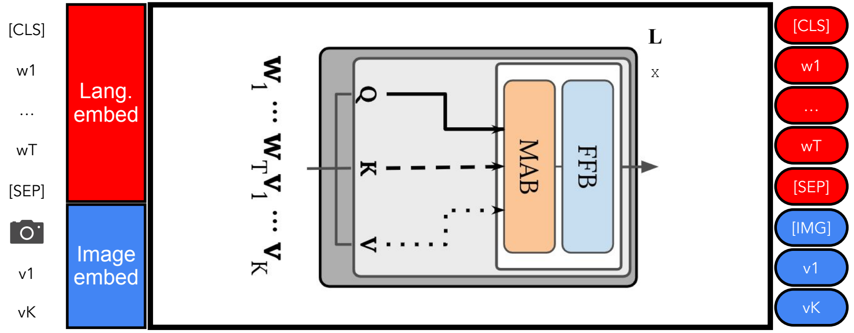
- Emerged as a key modelling approach in 2019 with a flurry of approaches to creating visually-grounded BERT models.
- This is a form of *model-based fusion*
- The backbone consists of two components:
    - language encoder
    - visual encoder

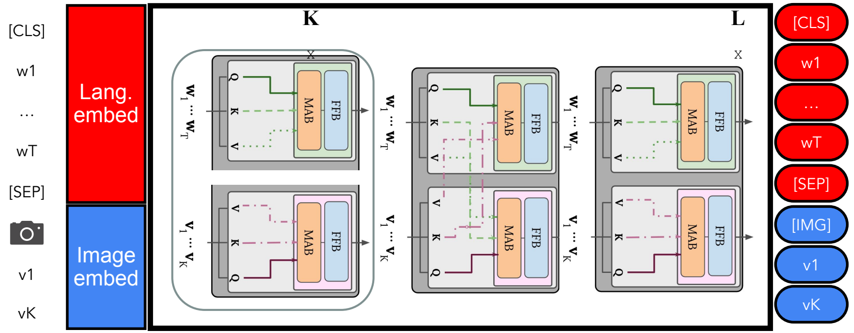


Tan & Bansal (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. EMNLP-IJCNLP.
Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Single- & Dual-Stream Architectures

- Single-stream
    - ○ Concatenate inputs into one sequence
- Dual-stream
    - ○ Process modalities independently
        - ■ Intra-modal
        - ■ Inter-modal

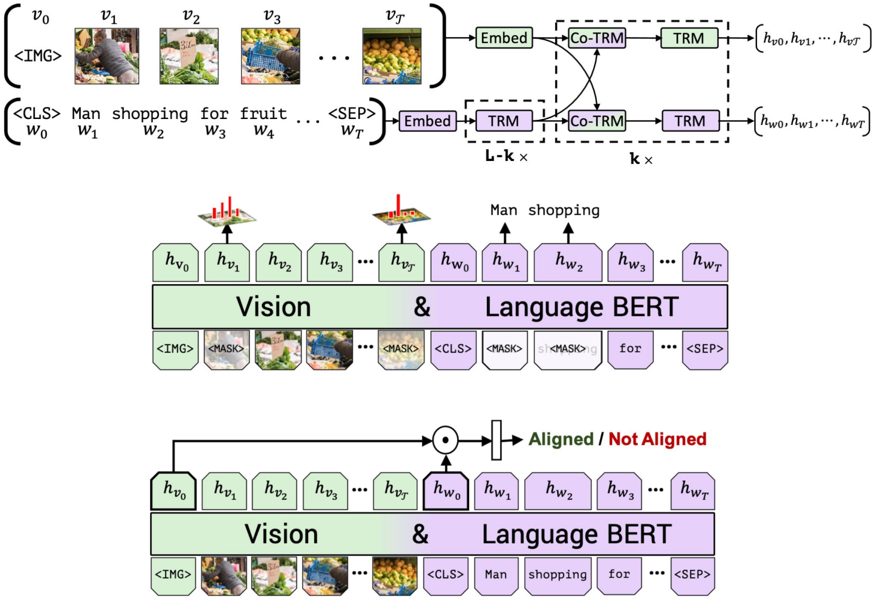




Bugliarello et al. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. TACL.

ViLBERT

- Dual-stream model
- Initialized from BERT
- Visual features extracted from 10-36 regions using Faster-RCNN
- Pretrained on Conceptual Captions
  - Masked Language Modelling
  - Masked Region Classification
  - Image-Text Matching

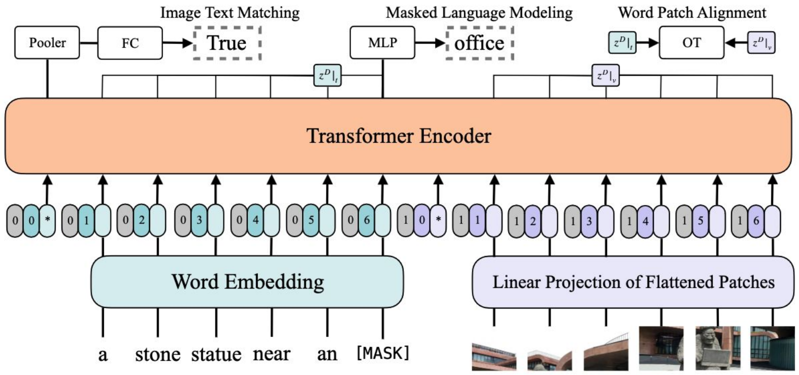


v_0
<IMG>
v_1
v_2
v_3
$v_\mathcal{T}$
<CLS> w_0
Man w_1
shopping w_2
for w_3
fruit w_4
<SEP> w_T
Embed
Co-TRM
TRM
L-k ×
k ×
$h_{v0}, h_{v1}, \cdots, h_{v\mathcal{T}}$
$h_{w0}, h_{w1}, \cdots, h_{wT}$
Man shopping
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... $h_{v_\mathcal{T}}$ h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <MASK> <MASK> <CLS> <MASK> <MASK> for <SEP>
Aligned / Not Aligned
<IMG> <CLS> Man shopping for <SEP>

Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

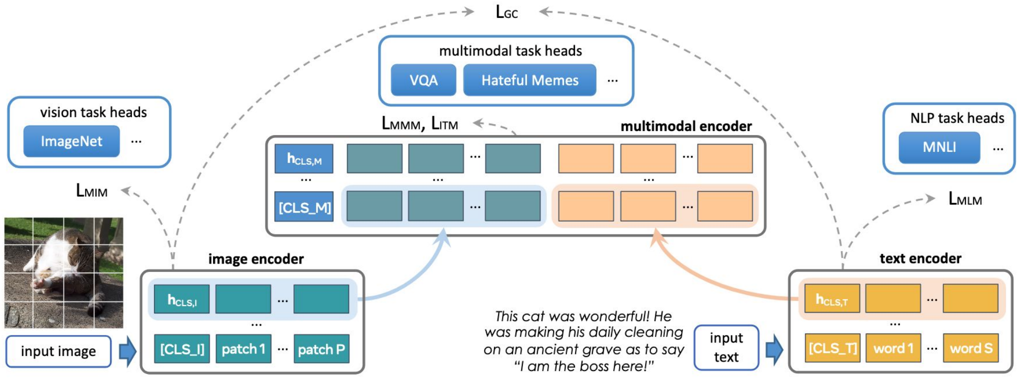
ViLT

- Single-stream model
- Initialized from BERT
- Visual features extracted from ViT-B/32
- Pretrained on Conceptual Captions, Visual Genome, COCO, SBU Captions
  - Masked Language Modelling
  - Image-Text Matching
  - Word-Patch Alignment



Kim et al. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. ICML.

FLAVA

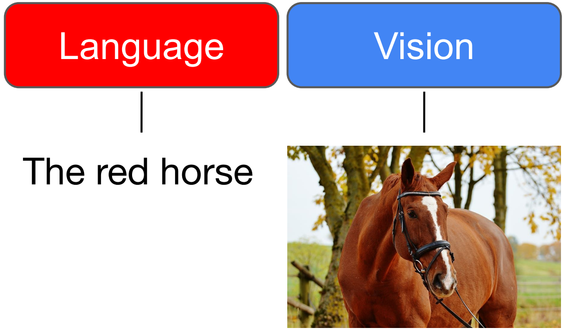


- Dual-stream Visual features extracted from ViT-B/16
- Pretrained on PMD70M
  - Masked Language Modelling, Masking Image Modelling
  - Image-Text Matching, Masked Multimodal Modelling
  - Global Contrastive Matching

Singh et al. (2022). FLAVA: A foundational language and vision alignment model. CVPR.

Dual-encoding Models

- Emerged as a sample-efficient alternative to cross-encoding.
- The backbone consists of two separate components:
  - language encoder
  - visual encoder



Radford et al. (2021). Learning transferable visual models from natural language supervision. ICML.

CLIP

Language

The red horse

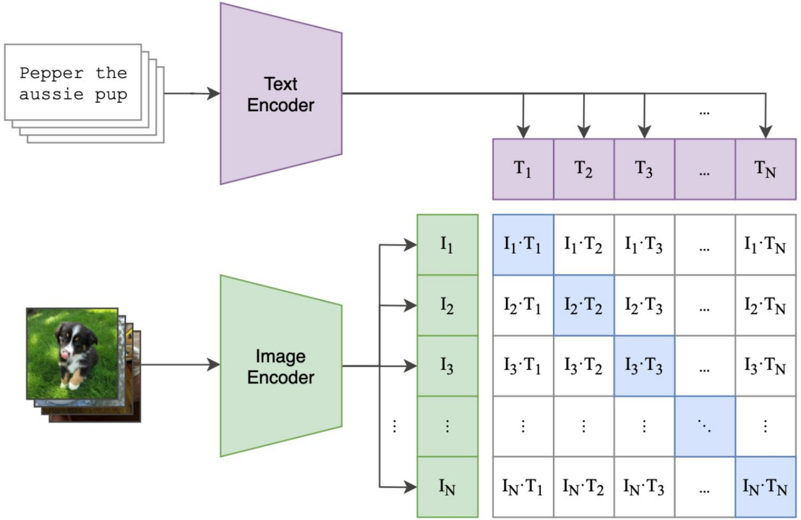
Vision

- 12 Layer Transformer Encoder
- ViT or ResNet Visual Encoder

- Maximize the similarity of the embeddings of paired examples (I, T):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{f(\mathbf{t},\mathbf{i})}{\sum_{\mathbf{t}'\in T} f(\mathbf{t}',\mathbf{i})}\right]$$

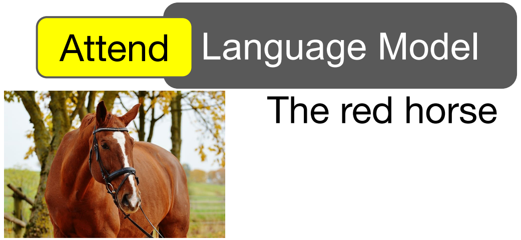
- Huge pretraining dataset of unclear provenance

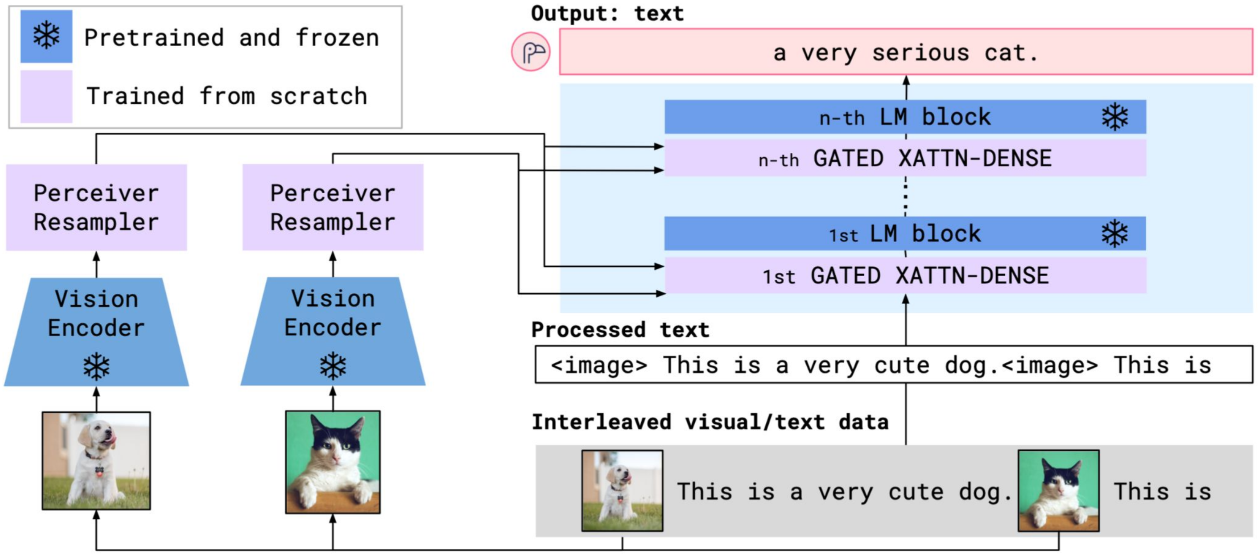


	T_1	T_2	T_3	...	T_N
I_1	$I_1{\cdot}T_1$	$I_1{\cdot}T_2$	$I_1{\cdot}T_3$	...	$I_1{\cdot}T_N$
I_2	$I_2{\cdot}T_1$	$I_2{\cdot}T_2$	$I_2{\cdot}T_3$	...	$I_2{\cdot}T_N$
I_3	$I_3{\cdot}T_1$	$I_3{\cdot}T_2$	$I_3{\cdot}T_3$	...	$I_3{\cdot}T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N{\cdot}T_1$	$I_N{\cdot}T_2$	$I_N{\cdot}T_3$	...	$I_N{\cdot}T_N$

Radford et al. (2021). Learning transferable visual models from natural language supervision. ICML.

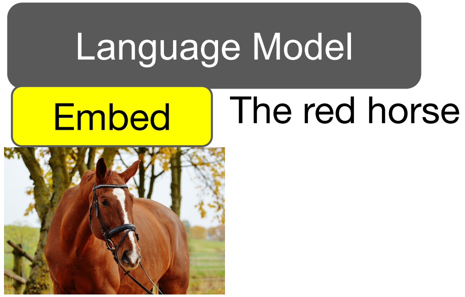
Cross-Attention



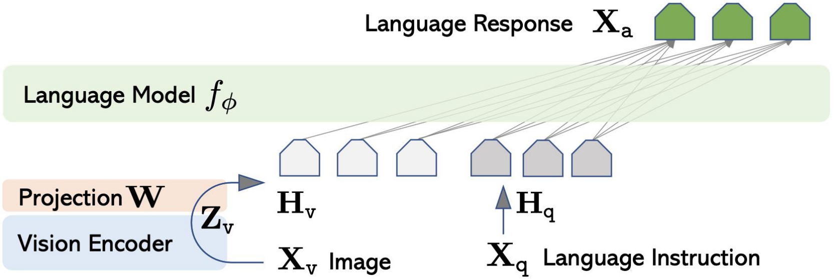


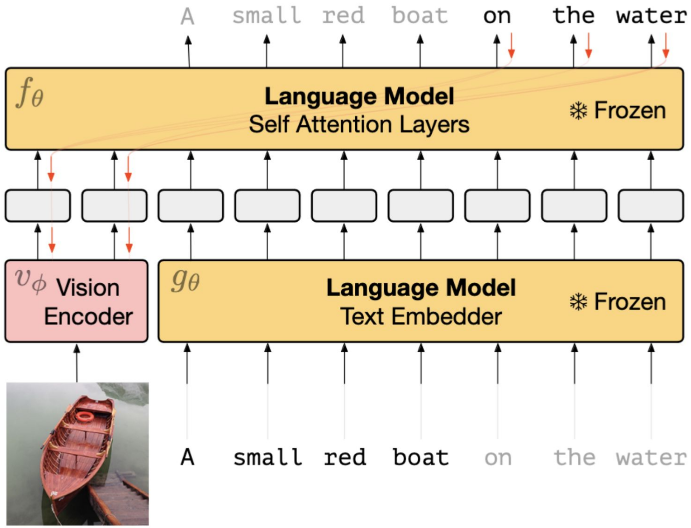
Alayrac et al. NeurIPS 2022. Flamingo: a Visual Language Model for Few-Shot Learning

Visual Prefix



- Exploit the representations learned during large-scale modality specific pretraining





Tsimpoukelli et al. NeurIPS 2021. Multimodal Few-Shot Learning with Frozen Language Models.
Liu et al. NeurIPS 2023. Visual Instruction Tuning.

Current Vision and Language Models

- Current models mostly follow this blueprint:

  1. Choose pretrained modality-specific components
  2. Learnable bridge between those components
  3. Dataset to estimate the parameters in the bridge
  4. Multi-stage finetuning strategy

Modality-Specific Components

	Vision Encoder	Language Model
LLAVA	CLIP ViT-L/14	Vicuna-13B
Qwen-VL	OpenCLIP ViT-bigG	Qwen-7B
MM1	ViT-L	1.3B LLM
PaliGemma	SigLIP-So400M/14	Gemma-2B

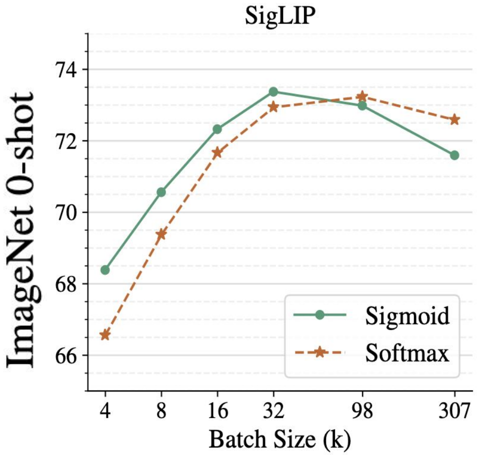
https://ai.google.dev/gemma/docs/paligemma
https://llava-vl.github.io/
https://github.com/QwenLM/Qwen-VL

SigLIP Image Encoder

“Unlike standard contrastive learning with softmax normalization, the sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization.”

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Label of the image-text pair: 1 if matched else -1

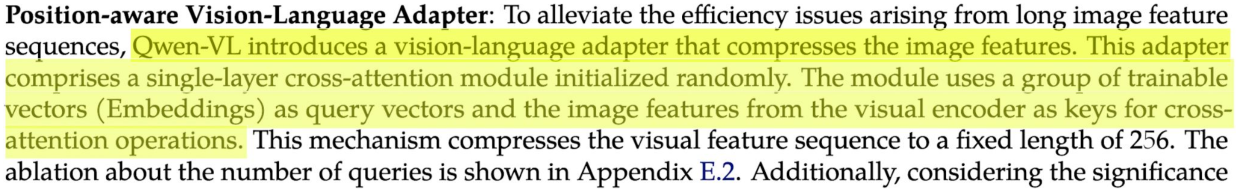


Zhai et al Sigmoid Loss for Language Image Pre-Training. ICCV 2023.

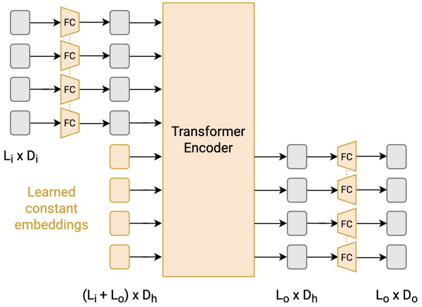
Learnable Bridge

- LLAVA and PaliGemma: Use a single linear layer to map the image output embeddings into the language model word embedding space.

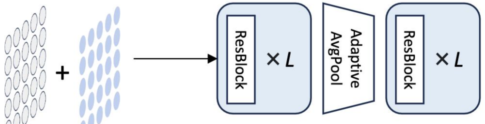
- Qwen-VL:



Position-aware Vision-Language Adapter: To alleviate the efficiency issues arising from long image feature sequences, Qwen-VL introduces a vision-language adapter that compresses the image features. This adapter comprises a single-layer cross-attention module initialized randomly. The module uses a group of trainable vectors (Embeddings) as query vectors and the image features from the visual encoder as keys for cross-attention operations. This mechanism compresses the visual feature sequence to a fixed length of 256. The ablation about the number of queries is shown in Appendix E.2. Additionally, considering the significance



- MM1: Convolutional-Abstractor
  - ResNet Block followed by an Adaptive Pooler



(a) C-Abstractor

Manas et al. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. EACL 2024.
Cha et al. Honeybee: Locality-enhanced Projector for Multimodal LLM. CVPR 2024

Training Dataset

- LLAVA: 595K image–caption examples filtered from CC3M

Qwen-VL	1.4 billion examples (77% English / 23% Chinese)
MM1	2+ billion mixture of image–text examples
PaliGemma	1 billion mixture of multilingual image caption, VQA, and in-the-wild datasets

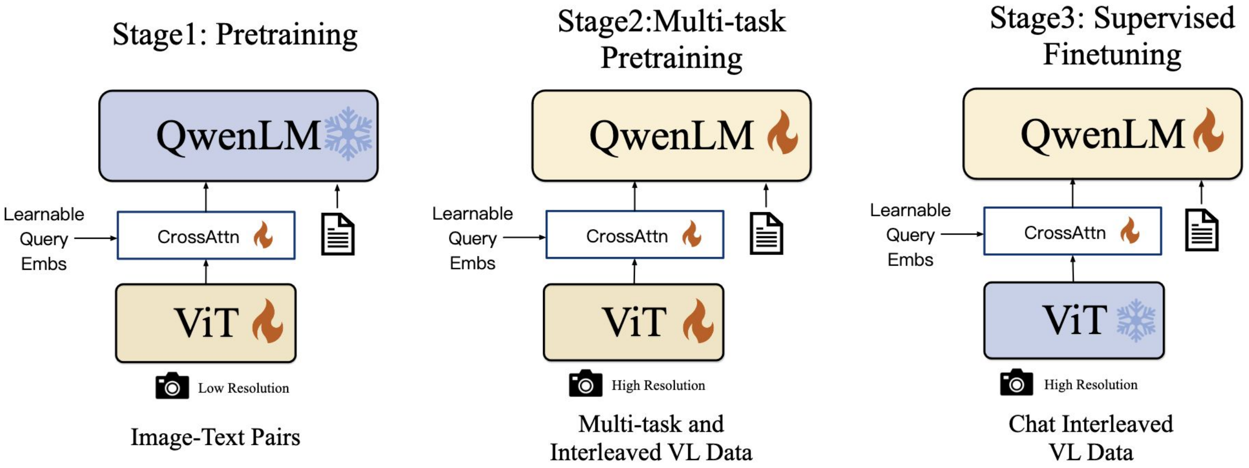
- The larger models are pretrained on **in-house data**
  - PaliGemma: WebLI (1B+), Qwen-VL (220M), MM1 (1B+)

Data Processing

- Encode the text using the language model tokenizer
- Encode the image using the image encoding model
- Image-position embeddings for multi-image sequences
- PaliGemma-specific
    - Location co-ordinate tokens
    - Segmentation tokens

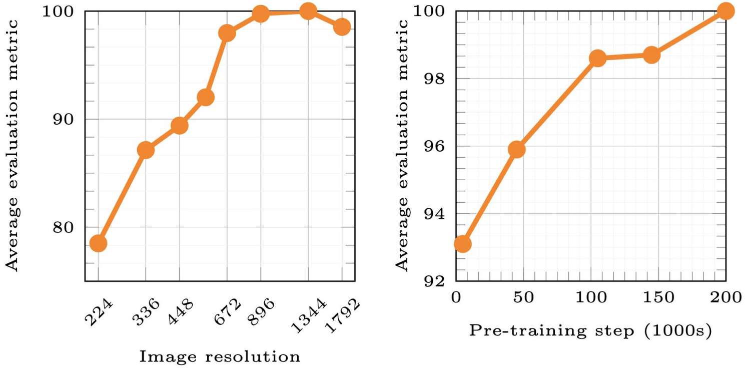
Training Strategy

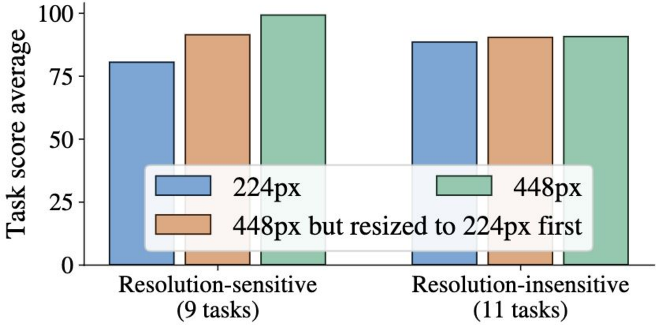
- Qwen-VL, PaliGemma, and MM1 use multi-stage training strategies with different types of data and different image resolutions



Open Questions

- How quickly will we realize these benefits in smaller models?
- Do LMMs really need 1 billion examples to learn a bridge?
- What happens to performance when we develop new tasks that involve weaker visual–linguistic bindings?





Summary

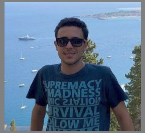
- Cross-encoding:
  - Many advances in which parts of the input contribute to loss
  - Shift from regions-of-interest to Vision Transformers
- Dual-encoding:
  - Excellent cross-domain transfer to a wide range of problems
- Visual Prefix Learning:
  - Exploit the benefits of single-modality pretraining

Q: Does an image captioning model need to learn everything in-weights?

PAELLA: Parameter-Efficient Lightweight Language-agnostic Captioning Model

Findings of NAACL 2024

R. Ramos



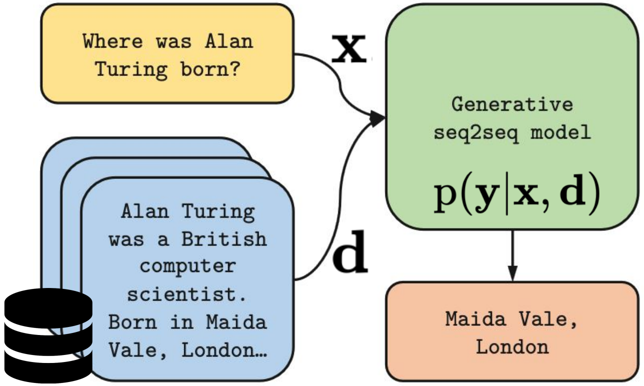
E. Bugliarello

B. Martins

D. Elliott

Retrieval Augmented Generation

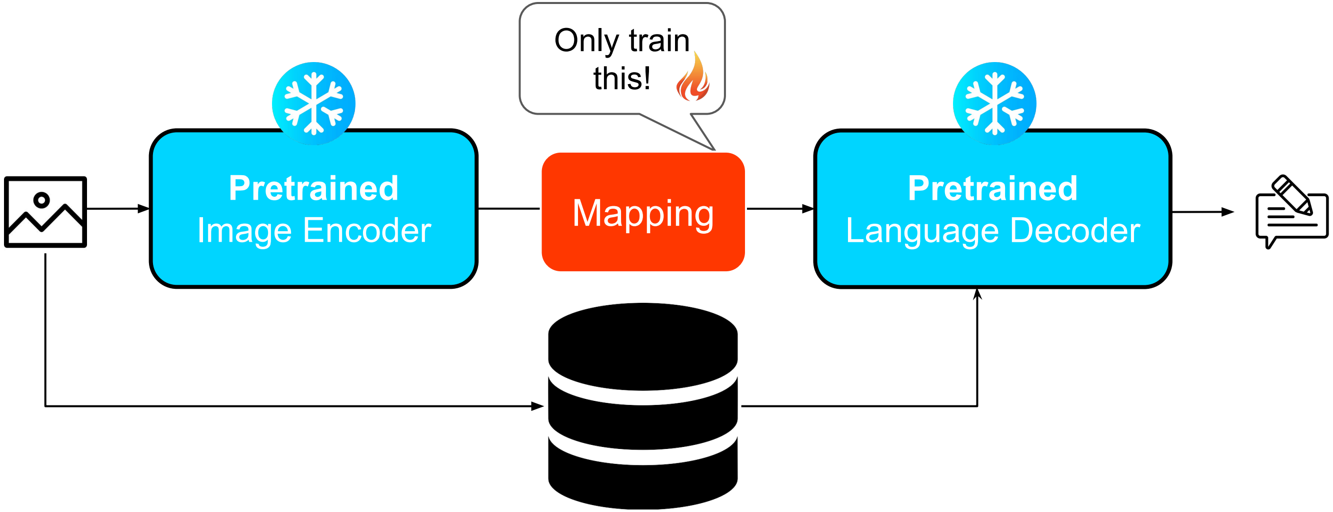
- Combine the power of in-weights learning with in-context adaptation through retrieval augmentation
- Given a datastore of facts, knowledge, documents, etc.
    - Combine the most relevant items from the datastore (d) with the input (x) for your task


Where was Alan Turing born?
x
Alan Turing was a British computer scientist. Born in Maida Vale, London…
d
Generative seq2seq model
$p(\mathbf{y}|\mathbf{x}, \mathbf{d})$
Maida Vale, London

Izacard and Grave. EACL 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering.

Motivation

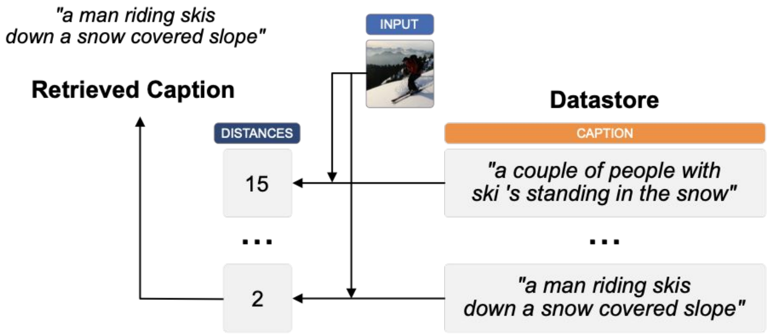
- Main trend in V&L is training bigger models on more data
- Alternative is emerging that re-uses independent backbone models
- Can we further improve performance with retrieval augmentation?



PAELLA Model

Retrieval System

- Build a FAISS datastore: store high-dimensional vectors
    - Captions of images represented with CLIP embeddings
- Retrieve k nearest-neighbours captions from datastore
    - Image embedding compared against datastore caption vectors


"a man riding skis down a snow covered slope"
Retrieved Caption
INPUT
DISTANCES
15
...
2
Datastore
CAPTION
"a couple of people with ski 's standing in the snow"
...
"a man riding skis down a snow covered slope"

Johnson et al. (2019). Billion-scale similarity search with GPUs. IEEE Big Data

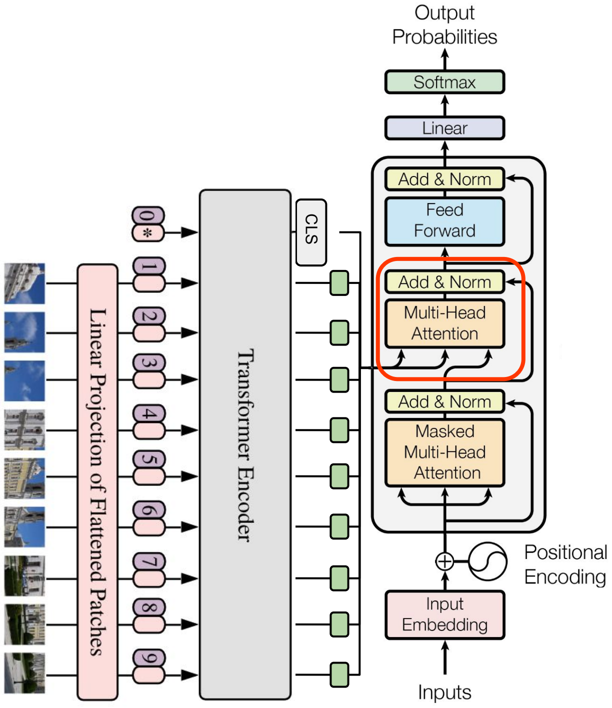
Trained Cross-Attention Layers

- We insert a randomly initialized **cross-attention mechanism** to attend to the visual encoder output embeddings

Rank	Params
d=128	553M
d=8	34M

$$\mathrm{Att}(\mathbf{X}\mathbf{W}^Q, \mathbf{X}\mathbf{W}^K, \mathbf{X}\mathbf{W}^V)$$

$$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{\text{d\_enc} \times d}$$



Experimental Protocol

- Encoder: Multilingual CLIP
- Decoder: XGLM-2.9B
- Training data:
  - 566K captions sampled uniformly from COCO-35
- Evaluation: XM-3600
  - 3600 geographically-diverse images
  - 36 languages: 100 captions per image
  - 5 low-resource languages (L5):
    - Bengali, Cusco Quechua, Maori, Swahili, Telugu

Examples images from XM3600

Results

	Data	Trained Θ	L36	L5
PaLI	12B	17B	53.6	-
Lg $_{\text{COCO-35}}$	19M	2.6B	15.0	12.5
mBLIP: BLOOMZ-7B	135M	800M	23.4	6.7
BB+CC$_{\textit{COCO-35 + CC-35}}$	135M	800M	28.5	22.4
mBLIP: mT0-XL	489M	124M	28.3	7.9
PAELLA	**566K**	**30M**	26.2	20.7

PAELLA is competitive against models with 35-863x more training data, and 4-87x more trained parameters

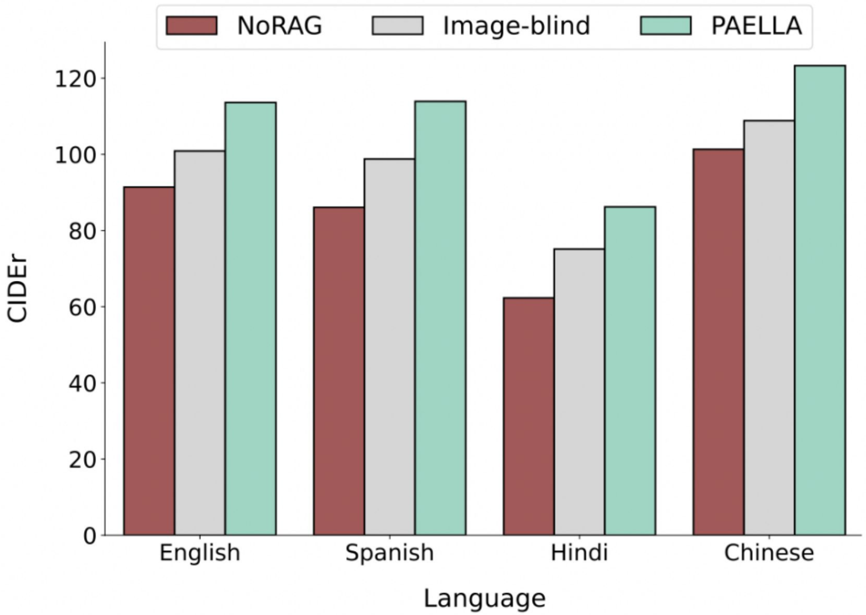
Zero-shot Multilingual Transfer

- **PAELLA**$_{\text{MONO}}$ is a variant trained on 566K examples in English COCO
- Outperforms **Lg** trained on 19.8M examples in the machine translated COCO-35 dataset

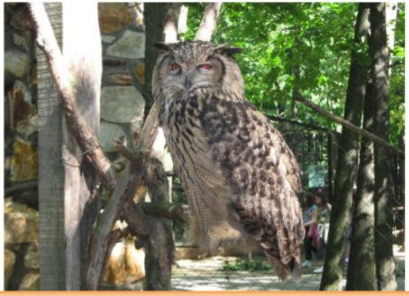
	Data	Trained ϴ	L36	L5
Lg: Thapliyal et al. $_{\text{COCO-35}}$	19M	2.6B	15.0	12.5
PAELLA$_{\text{MONO}}$	566K$_{\text{en}}$	30M	15.5	12.1

Value of Retrieval Augmentation

Consistent improvements from multilingual retrieval augmentation across the core languages in the XM3600 evaluation data



Qualitative Example



类似图片显示：

ऐसी ही तस्वीरें दिखाती हैं:

Imágenes similares muestran:

Similar images show:

the owl is perched outside in front of the people
an owl sitting a top a table during the daytime
an owl is sitting on a perch at a camp site
the fuzzy owl is sitting on a tree branch

A caption I can generate to describe this image in **english** is:

PAELLA

en: *"an owl sitting on top of a tree"*

es: *"un búho sentado en una rama de un árbol"*
(an owl sitting on a tree branch)

hi: *"एक उल्लू एक पेड़ की टहनी पर बैठा है"*
(an owl is sitting on a tree branch)

zh: *"一只 猫头鹰 站在 树上"*
(an owl standing in a tree)

NoRAG

en: *"a large black and white picture of a bird"*

es: *"un pájaro posado en la parte superior de un edificio"*
(a bird perched on the top of a building)

hi: *"एक पेड़ के पास खड़ा एक पक्षी"*
(a bird standing near a tree)

zh: *"一只 长颈鹿 坐在 树枝 上"*
(a giraffe sitting on a branch)

Try it yourself (in English)

Demo: https://huggingface.co/spaces/RitaParadaRamos/SmallCapDemo

Sequential Compositional Generalization in Multimodal Models

NAACL 2024

S. Yagcioglu

O. B. Ince

A. Erdem

E.
Erdem

D. Elliott

D. Yuret

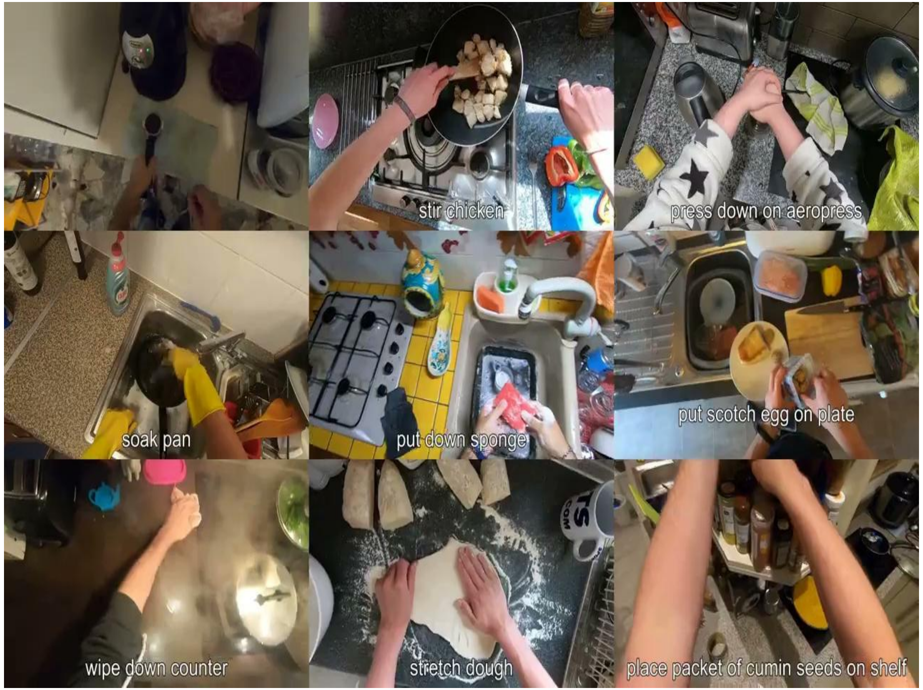
Why Compositionality?

- Given recent advances in MLLMs, we should work on tasks that require more sophisticated logical or commonsense reasoning
    - RecipeQA (Yagcioglu et al. 2018)
    - ScienceQA (Lu et al. 2022)
- **Sequential Multimodal Compositional Generalization** requires models to reason across a sequence of related multimodal inputs

RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. Yagcioglu et al. EMNLP 2018.
Learn to explain: Multimodal reasoning via thought chains for science question answering. Lu et al. NeurIPS 2022.

EPIC Kitchens 100

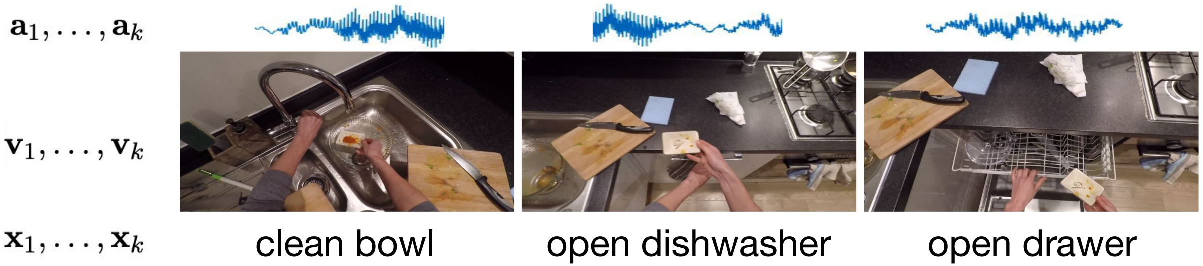
- Head-mounted recordings from 45 different kitchens
- Rich annotations of objects
- Simple action descriptions
  - 93 verb classes
  - 300 object classes



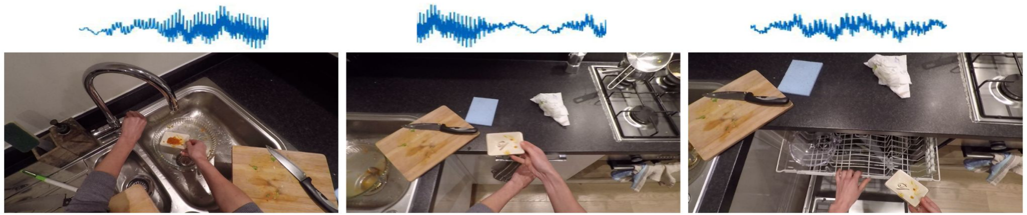
Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. Damen et al. IJCV 2022.

CompAct Dataset

- Multimodal sequences consisting of aligned segments
  - Audio recording (not speech)
  - Video frames
  - Short description


$\mathbf{a}_1, \ldots, \mathbf{a}_k$
$\mathbf{v}_1, \ldots, \mathbf{v}_k$
$\mathbf{x}_1, \ldots, \mathbf{x}_k$
clean bowl
open dishwasher
open drawer

Compositional Generalization Tasks



clean bowl open dishwasher open drawer

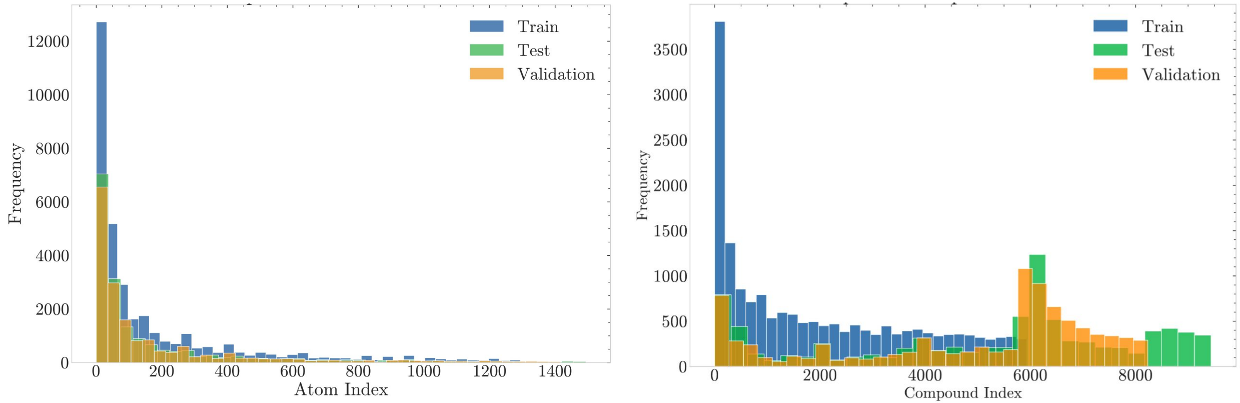
1. Next Description Prediction using a language model

$$p_\theta(\mathbf{y} = \mathbf{x}_{k+1} | \mathbf{x}_{1:k}, \mathbf{v}_{1:k}, \mathbf{a}_{1:k})$$

2. Atom Classification
    - Verb $p_\theta(\mathbf{y} = v | \mathbf{x}_{1:k}, \mathbf{v}_{1:k}, \mathbf{a}_{1:k})$
    - Object $p_\theta(\mathbf{y} = o | \mathbf{x}_{1:k}, \mathbf{v}_{1:k}, \mathbf{a}_{1:k})$

Atoms and Compositionality

- We use Maximum Compound Divergence (Keysers et al. 2020) to create a dataset that requires compositional generalization
  - ○ Noun and verbs extracted from simple descriptions

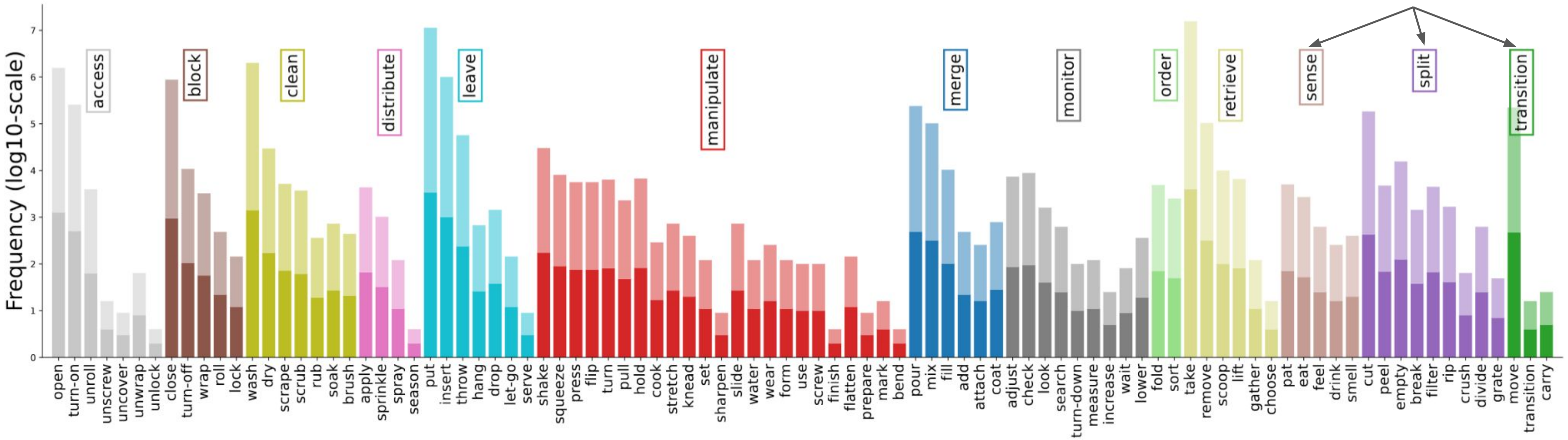


Measuring compositional generalization: A comprehensive method on realistic data. Keysers et al. ICLR 2020.

CompAct Dataset Statistics

- 8,766 multimodal sequences
  - 50% training, 25% validation, 25% test

category



Distribution of verb classes in CompAct

Models

Baselines

Trained on CompAct:

- **T**ext-only
- **V**ision & **L**anguage
- **O**bject & **L**anguage
- **A**udio & **L**anguage
- **V**ision, **A**udio, & **L**anguage
- **O**bject, **A**udio, & **L**anguage

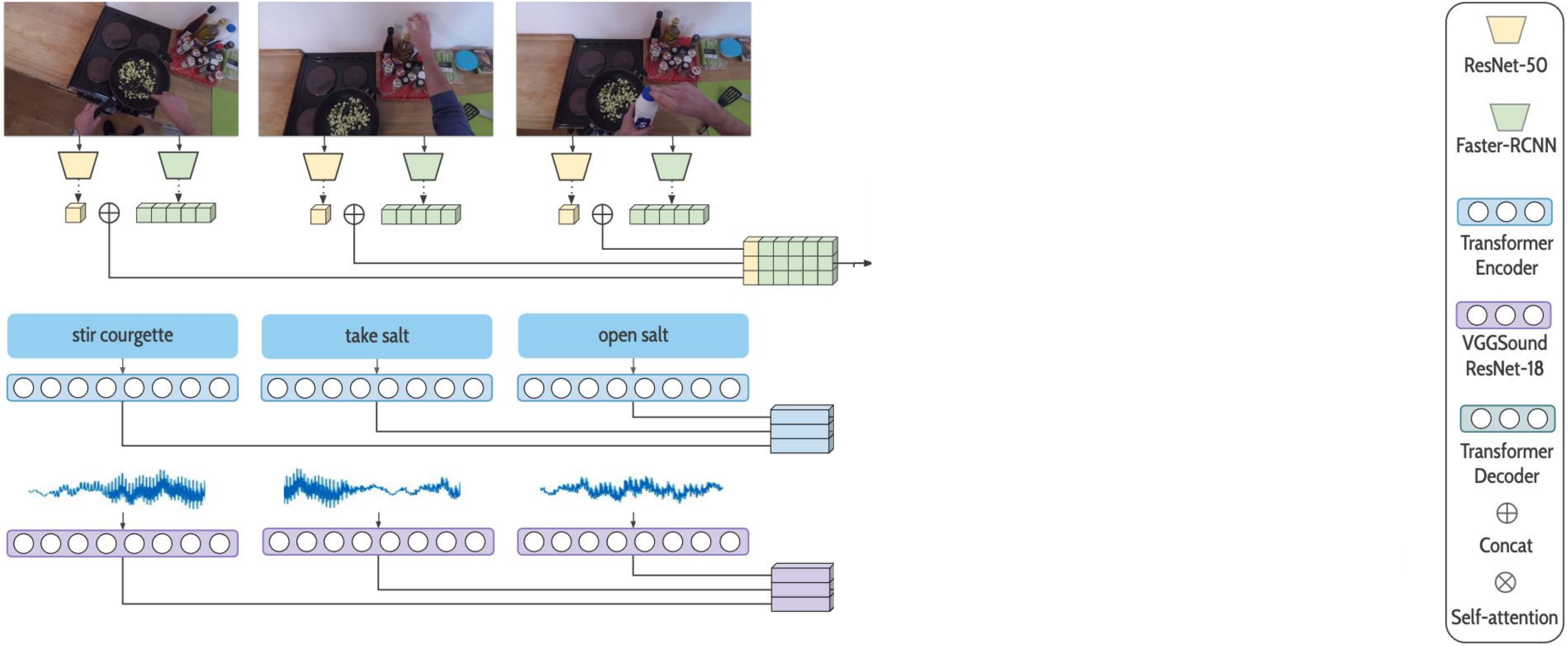
Pretrained (M)LLMs

k=8-shot prompting:

- LLaMa-2 Chat 7B
- IDEFICS-9B
- OpenFlamingo-9B
- Otter-7B

No guarantee that these models need to compositionally generalize

Baseline Architecture



Example Prompt Templates

LLaMA-2 Prompt (k=3)

```
Predict the next narration given 3 sequential previous narrations from a cooking video
put down bowl . move frying pan . pick up spatula => put down spatula
move yoghurt . put down bowl . pick up yogurt => put yoghurt
put down bowl . grab wok . move tap => lather wok
pick up tins . put down tins . move bowl =>
```

IDEFICS Prompt (k=1)

```
Predict the next action narration given 3 sequential previous actions (image-narration pairs) in a cooking video.
put down bowl <Image 1> . move frying pan <Image 2> . pick up spatula <Image 3> => put down spatula
pick up tins <Image 1> . put down tins <Image 2> . move bowl <Image 3> =>
```

Experimental Details

- Video segments represented by one keyframe
  - Based on detected objects from Faster R-CNN

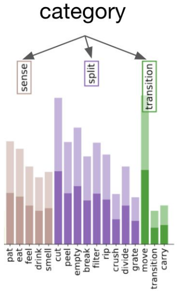
- Baselines tokenize multi-part tokens as one token
  - olive oil -> olive_oil

- MLLMs use own tokenizers

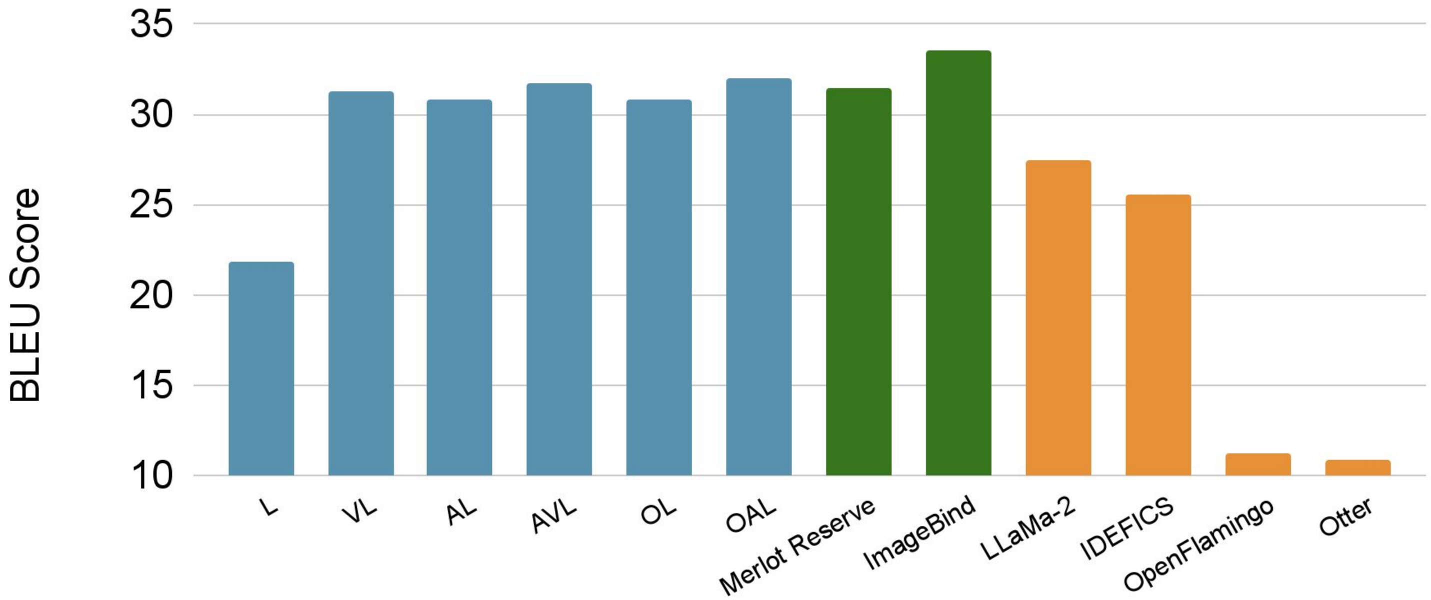
- All experiments used one 16GB NVIDIA V100.
  Each baseline can be trained in < 1 hour

Evaluation Measures

- BLEU score
- Exact Match
  - How often does the model predict exactly the expected noun or verb?
- Categorical Accuracy
  - Does the model predict a noun or verb in the same semantic category?



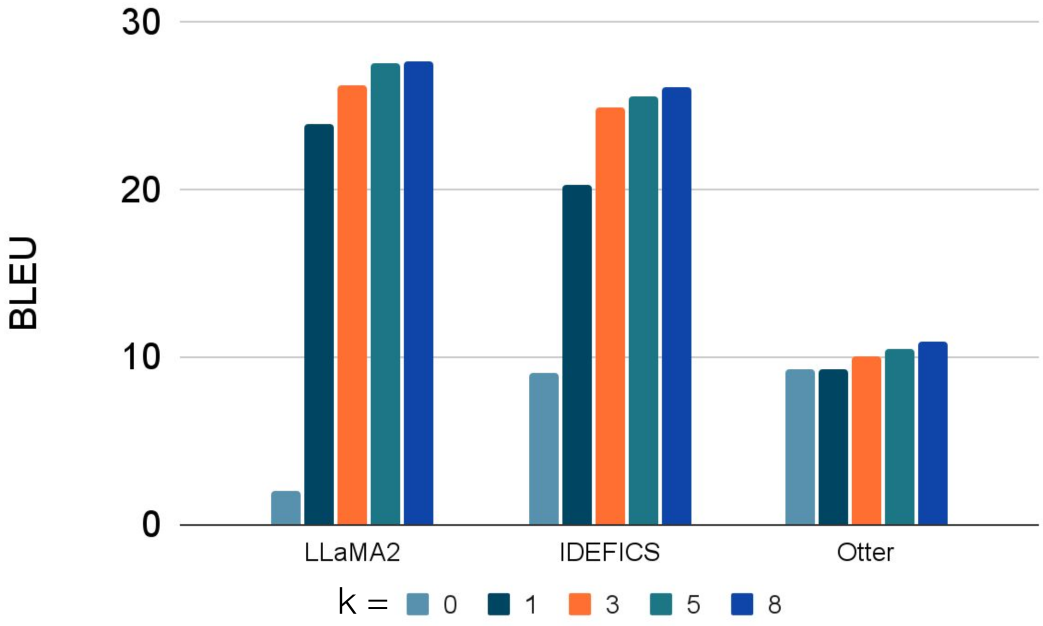
Next Utterance Prediction



Better features are useful

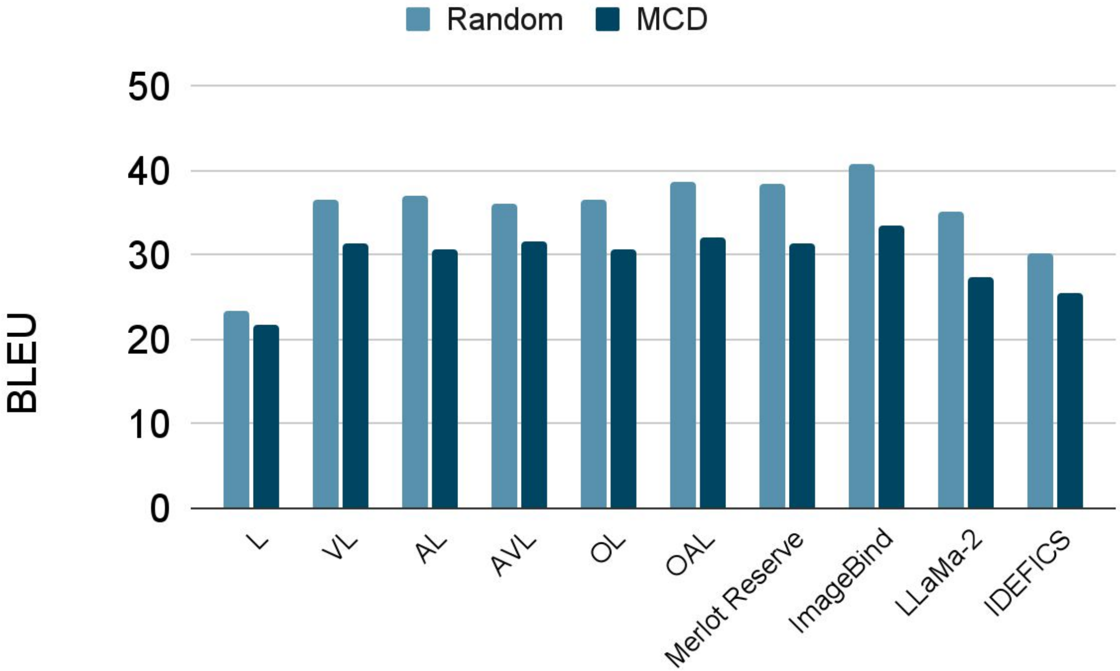
Few-shot prompted MLLMs fall short

Few-shot Prompting



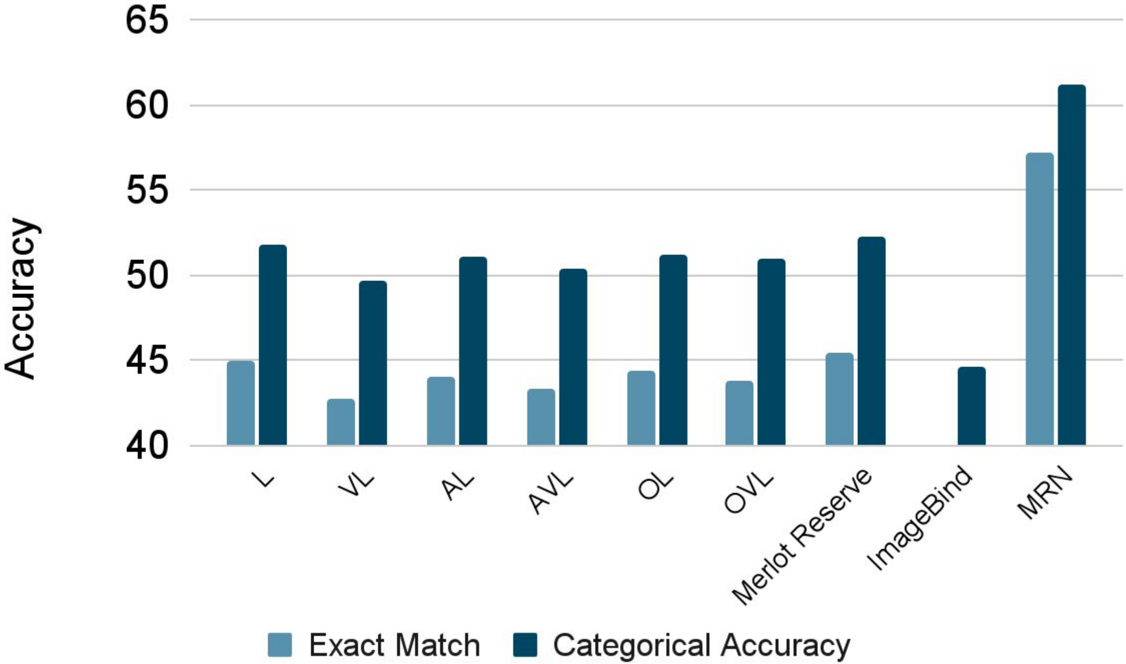
Clear benefit from increasing the number of in-context examples

Random I.I.D. split data is easier



All models generalize better to randomly split data

Noun Prediction



No model can outperform simple most recent noun heuristic

Qualitative Examples

clean bowl

open dishwasher

open drawer

L: close dishwasher
AL: close dishwasher
VL: close dishwasher
AVL: dry bowl
IB: put bowl in dishwasher
LLaMA-2: get clean
IDEFICS: put bowl away

put pan in drainer

pick up mug

pick up sponge

L: put sponge
AL: sponge mug
VL: sponge mug
AVL: sponge mug
IB: sponge mug
LLaMA-2: put sponge in sink
IDEFICS: sponge mug

Conclusions

- Sequential Multimodal Compositional Generalization is challenging new task where better multimodal features improve performance
- We find no evidence that ICL or large-scale multimodal pretraining can solve this task
- Future work includes
  - integrating even better features into the baseline
  - fine-tuning MLLMs using LoRA
  - including more keyframes in the visual input

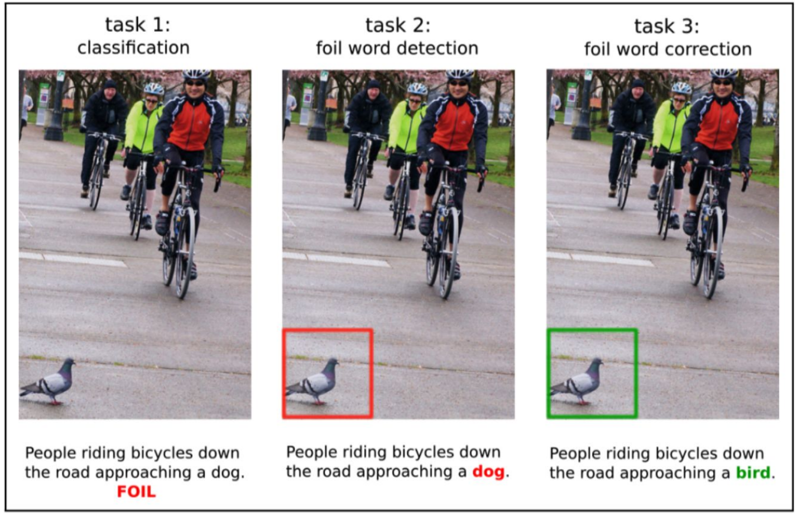
4. Understanding Multimodal Models

Beyond Benchmarking

- Many questions about what drives the success of these models?
  - Better contextualization: make better use of the multimodal inputs
  - Acquire certain “skills”, e.g. counting or localization
  - Understand linguistic structures
  - Something else?
- Model-internal behaviour
  - Attention mechanism patterns
- Probing
  - Tasks related to different skills

FOIL Captions

- Do V&L models really understand the relationship between words and images?
- Crowdsource datasets that contain contextually plausible but incorrect image–text pairs


task 1:
classification
task 2:
foil word detection
task 3:
foil word correction
People riding bicycles down the road approaching a dog.
FOIL
People riding bicycles down the road approaching a dog.
People riding bicycles down the road approaching a bird.

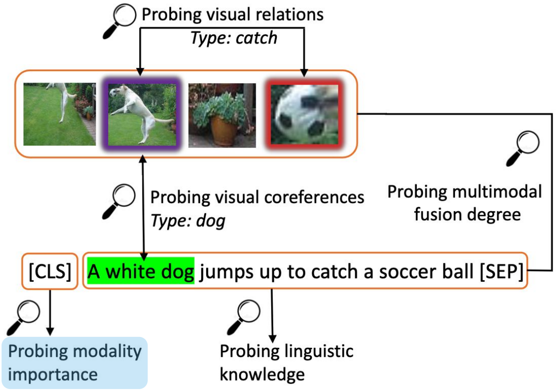
Shekhar et al. (2017). FOIL it! Find One mismatch between Image and Language caption. ACL.

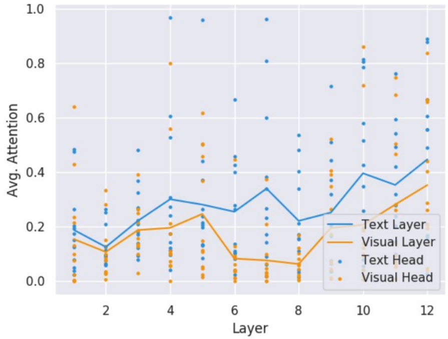
Vision and Language Understanding Evaluation

- Suite of five model probing tasks
- **Modality Influence**: Estimate the layer-wise contribution of each modality to the `[CLS]` embedding:

$$I_{M,j} = \sum_{i \in S} \mathbb{1}(i \in M) \cdot \alpha_{ij}$$

  - The UNITER model relies more on textual features when fusing modalities throughout the model


Probing visual relations
Type: catch
Probing visual coreferences
Type: dog
Probing multimodal fusion degree
[CLS]
A white dog jumps up to catch a soccer ball [SEP]
Probing modality importance
Probing linguistic knowledge

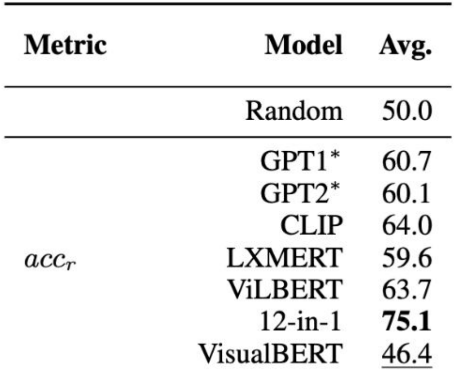


Cao et al. (2020). Behind the scene: Revealing the secrets of pre-trained vision-and-language models. ECCV.

VALSE Benchmark

- Test visio-linguistic capabilities with image-sentence foil pairs
- Image-sentence matching task
  - Existential quantifiers
  - Semantic number
  - Counting
  - Prepositional relations
  - Action replacement / swap
  - Co-reference

A small copper vase with some flowers / exactly one flower in it.



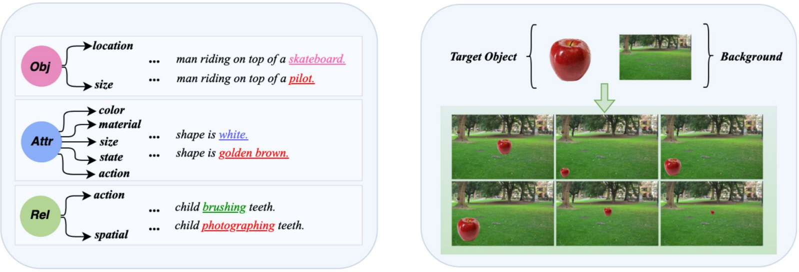
Metric	Model	Avg.
	Random	50.0
acc_r	GPT1*	60.7
	GPT2*	60.1
	CLIP	64.0
	LXMERT	59.6
	ViLBERT	63.7
	12-in-1	**75.1**
	VisualBERT	46.4

$p(caption, img) > p(foil, img)$

Parcalabescu et al. (2022). VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena. ACL.

VL-CheckList

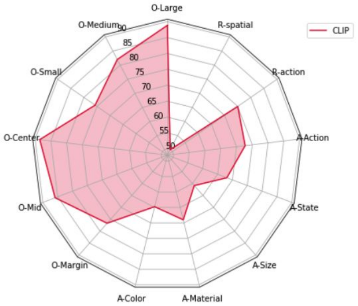
- Evaluate V&L models based on automatic manipulations to vision and language data.
- Image-Sentence matching task
- Radar chart overviews based on object / attribute / relationship variations



Obj
location
size
... man riding on top of a skateboard.
... man riding on top of a pilot.
Attr
color
material
size
state
action
... shape is white.
... shape is golden brown.
Rel
action
spatial
... child brushing teeth.
... child photographing teeth.
Target Object
Background

(a) Language Variation

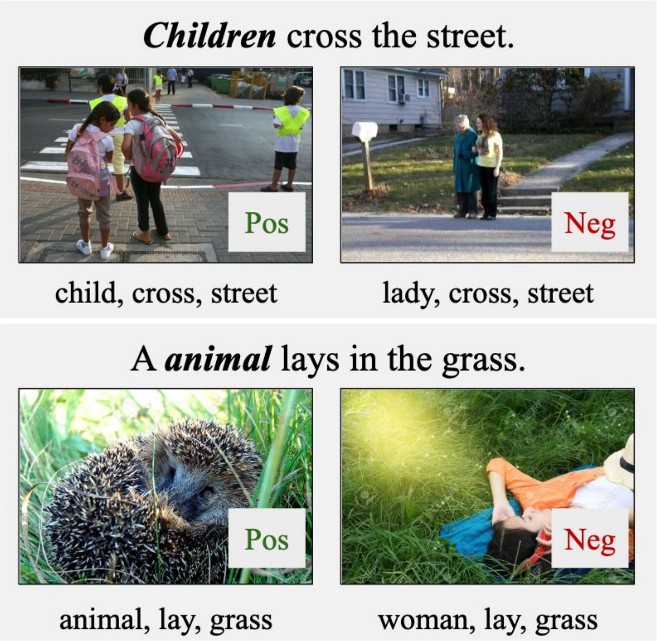
(b) Vision Variation



Zhao et al. (2022). An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models. EMNLP.

Subject-Verb-Object Probes

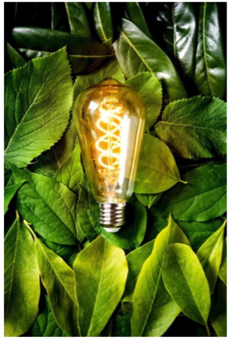
- Large-scale dataset with SVO triplets mined from Conceptual Captions and 14K images and with crowdsourced captions
- Foil detection formulation



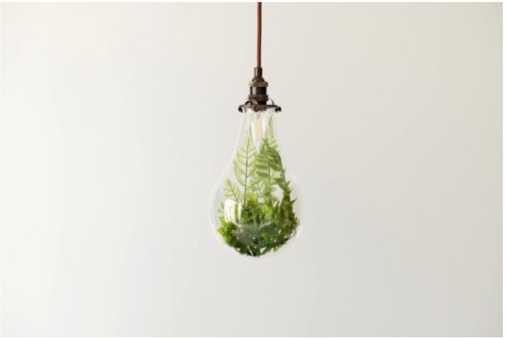
Hendricks and Nematzadeh. (2021). Probing Image–Language Transformers for Verb Understanding. ACL.

WinoGround

- 1,600 text-image pairs to evaluate compositional understanding



some plants surrounding a lightbulb

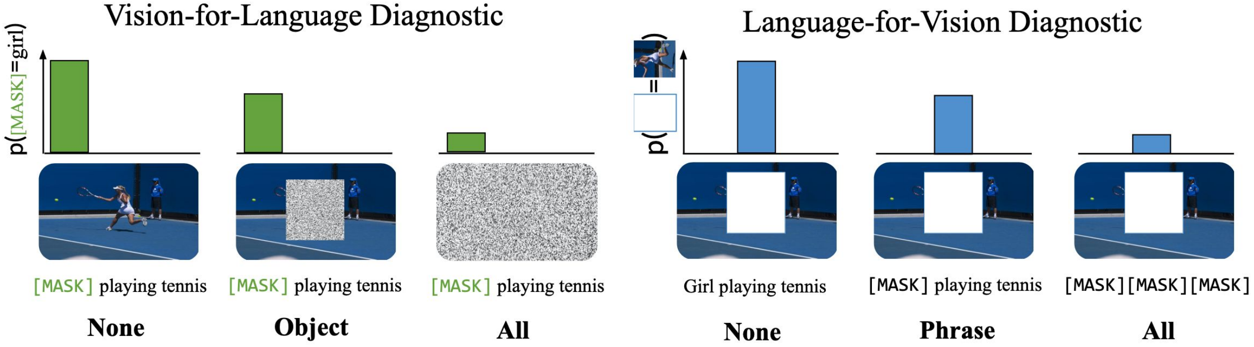


a lightbulb surrounding some plants

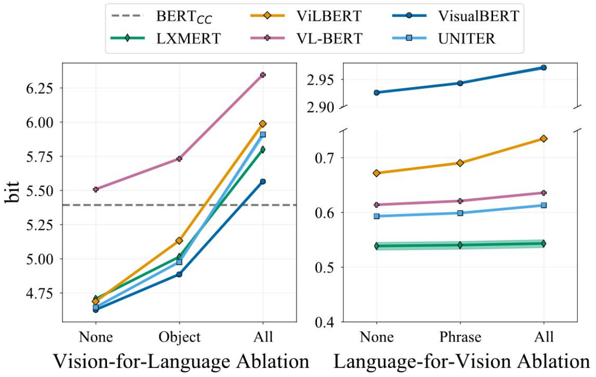
- Images sourced **with permission** from Getty.
- Differences are categorised into: swap dependent, swap-independent, and visual differences

Thrush et al. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. CVPR.

Vision-for-Language?



- Pair of diagnostic evaluations that can be applied to any model that makes MLM and MRC predictions.



Frank et al. (2021). Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. EMNLP.

Summary

- Understanding and analysis is a vibrant area of research
- Foil detection is the most popular methodology
- Witnessing a methodological shift
  - attention analyses → linguistically-informed analyses
  - hand-crafted datasets
  - simpler accuracy-based metrics

5. Future Directions

Physical Understanding

- Predicting and explaining physical actions in the world will become of increasing importance as we create embodied agents

Q: How many objects are prevented by the tiny green triangle from falling into the basket?

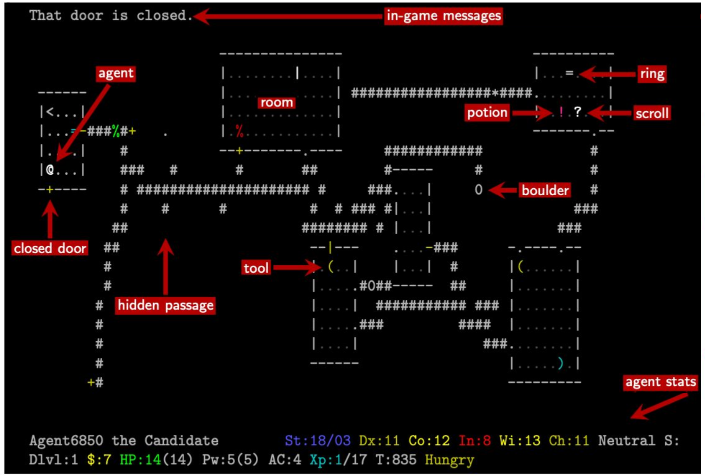
Q: What is the color of the last object that collided with the tiny red circle?

Q: If any of the other objects are removed, will the tiny green circle end up in the basket?

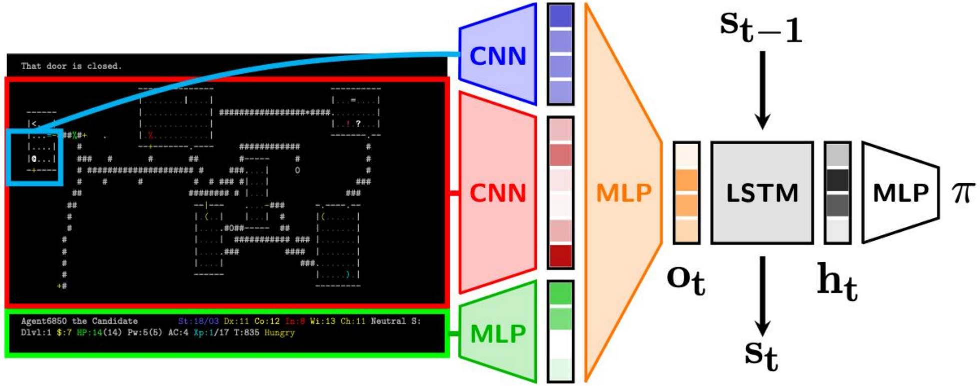
Ates et al. (2022). CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions. ACL.

Text-based Video Games

- Learning to act in procedurally-generated video game environments with rich contexts, action spaces, and long-term rewards



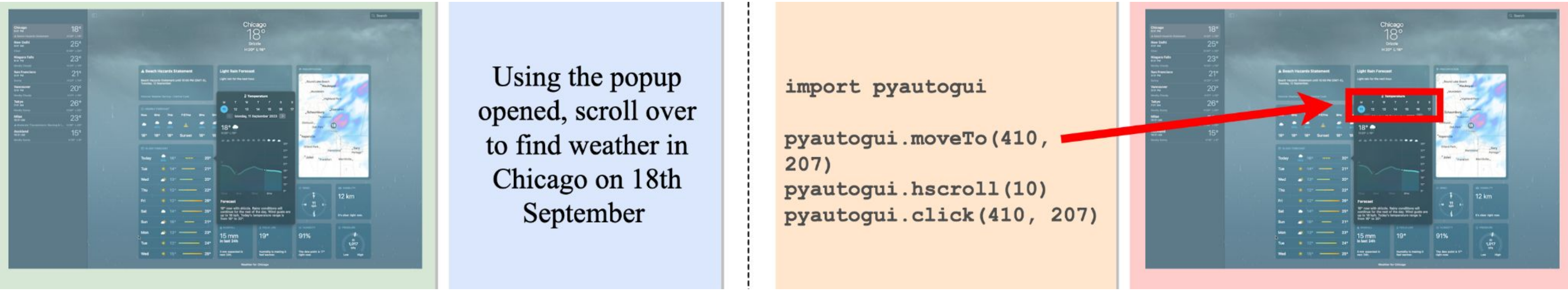
```
That door is closed.

Agent6850 the Candidate          St:18/03 Dx:11 Co:12 In:8 Wi:13 Ch:11 Neutral S:
Dlvl:1 $:7 HP:14(14) Pw:5(5) AC:4 Xp:1/17 T:835 Hungry
```



Küttler et al. (2020). The NetHack Learning Environment. NeurIPS.

Multimodal Interaction

- OmniAct combines multimodal understand with program execution to solve a variety of tasks that humans perform with their computers



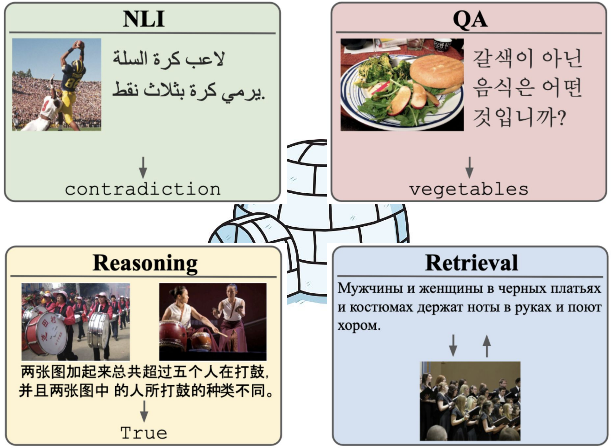
Using the popup opened, scroll over to find weather in Chicago on 18th September

```
import pyautogui

pyautogui.moveTo(410,
207)
pyautogui.hscroll(10)
pyautogui.click(410, 207)
```

Kapoor et al. ECCV 2024. A Dataset and Benchmark for Enabling Multimodal Generalist Autonomous Agents for Desktop and Web.

Multilinguality

- The majority of Vision and Language research is in English
- We need resources, models, and evaluations to create useful multilingual multimodal models
- High-quality data requires:
    - time
    - money
    - community engagement



لاعب كرة السلة
يرمي كرة بثلاث نقط.

갈색이 아닌
음식은 어떤
것입니까?

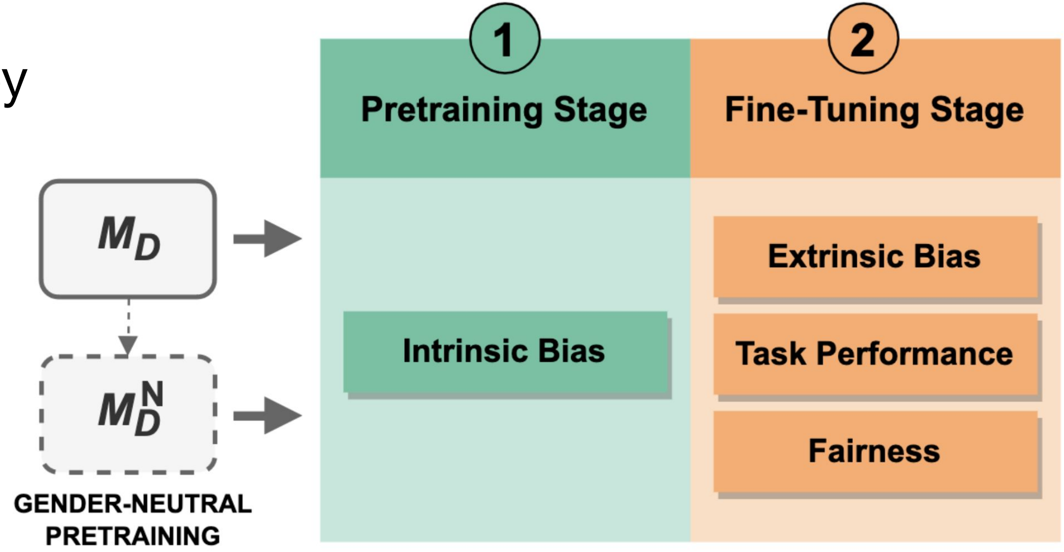
两张图加起来总共超过五个人在打鼓，
并且两张图中 的人所打鼓的种类不同。

Мужчины и женщины в черных платьях и костюмах держат ноты в руках и поют хором.

Bugliarello et al. ICML 2022. IGLUE: Image-Grounded Language Understanding Evaluation.

Bias and Fairness

- What are the intrinsic biases learned during multimodal pretraining and how do they affect downstream task performance?

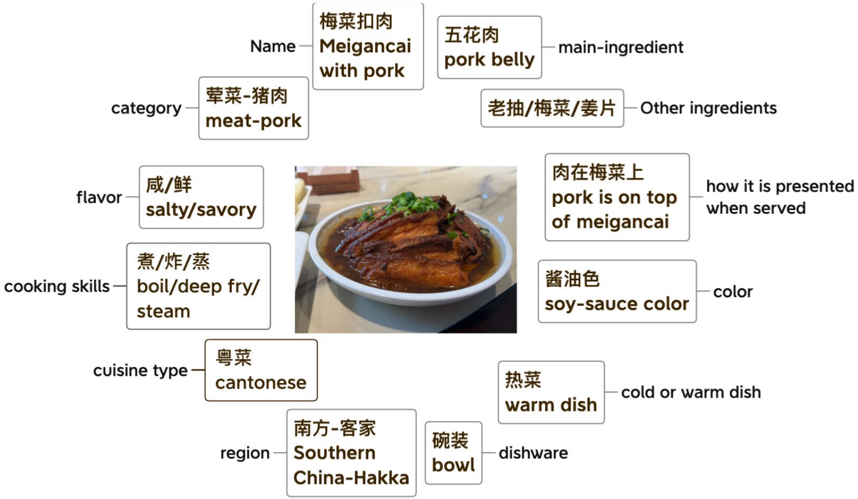


Cabello et al. EMNLP 2023. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models

Next

Fine-grained Multimodal Data ...

- Domain-specific fine-grained VQA data
  - Chinese food
  - Highly-detailed taxonomy
  - Human questions
  - Three version of the task
  - Private data



Name — 梅菜扣肉 Meigancai with pork

五花肉 pork belly — main-ingredient

category — 荤菜-猪肉 meat-pork

老抽/梅菜/姜片 — Other ingredients

flavor — 咸/鲜 salty/savory

肉在梅菜上 pork is on top of meigancai — how it is presented when served

cooking skills — 煮/炸/蒸 boil/deep fry/ steam

酱油色 soy-sauce color — color

cuisine type — 粤菜 cantonese

热菜 warm dish — cold or warm dish

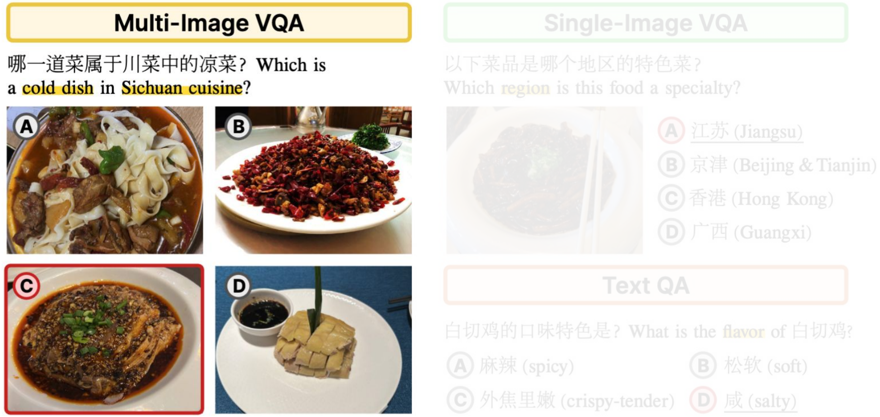
region — 南方-客家 Southern China-Hakka

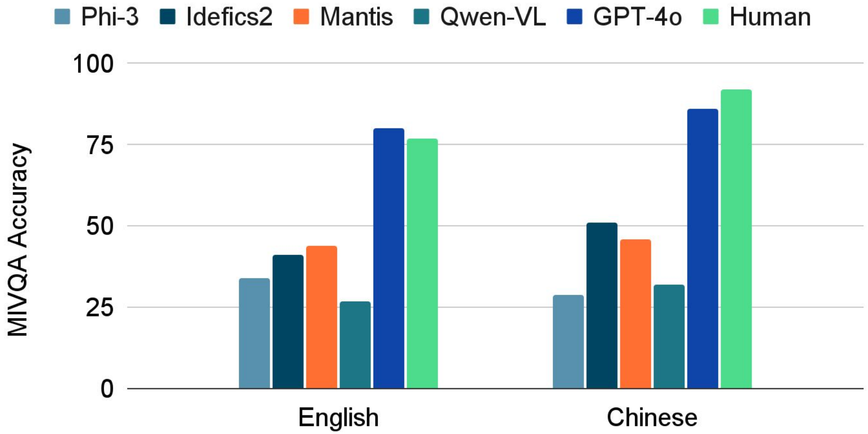
碗装 bowl — dishware

Li et al. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture. 2024.

Next

is challenging for LMMs


Multi-Image VQA
哪一道菜属于川菜中的凉菜？Which is a cold dish in Sichuan cuisine?
A
B
C
D
Single-Image VQA
以下菜品是哪个地区的特色菜？Which region is this food a specialty?
A 江苏 (Jiangsu)
B 京津 (Beijing & Tianjin)
C 香港 (Hong Kong)
D 广西 (Guangxi)
Text QA
白切鸡的口味特色是？What is the flavor of 白切鸡?
A 麻辣 (spicy)
B 松软 (soft)
C 外焦里嫩 (crispy-tender)
D 咸 (salty)



API-based model barely outperforms a naive English annotator

Huge gap to fill between API-based and open-weights models

Next wonders: Community-driven Multimodal Data

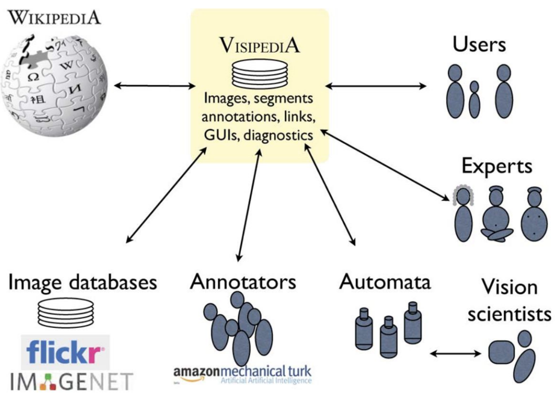
- Large-scale data collection with crowdworkers: most budgets cannot scale

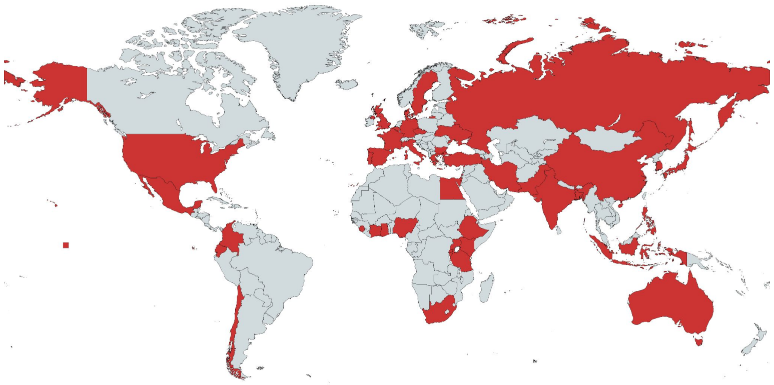
Core team
- Define data & task
- Recruit and coordinate Researchers (eg PhDs)

Researchers
- Coordinate annotation
- Verify annotations

Crowdworkers
- Create annotations

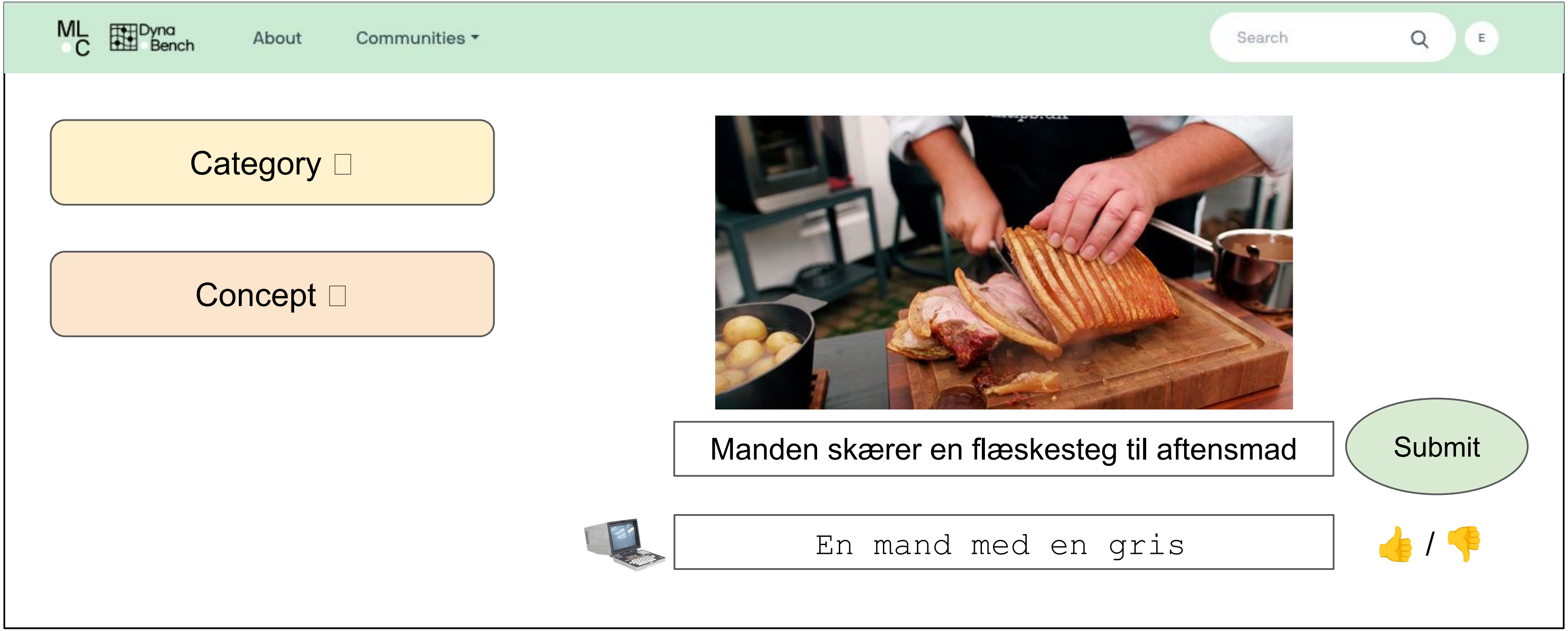
- Our bet: **People care about their culture**




Created with mapchart.net

Next

Gamified Data Collection



Manden skærer en flæskesteg til aftensmad

En mand med en gris

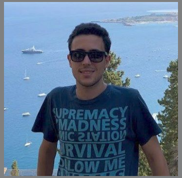
Q: What if we treated language as vision?

Language Modelling with Pixels

ICLR 2023

P. Rust

J. F. Lotz



E. Bugliarello

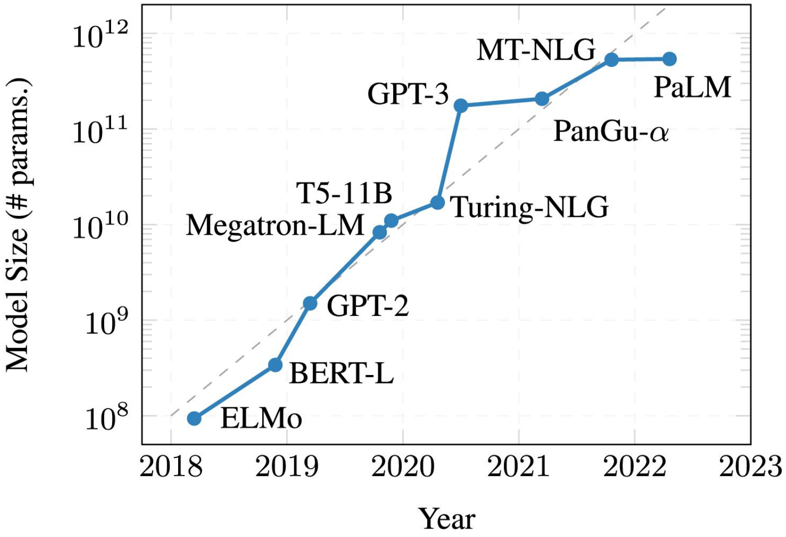
E. Salesky

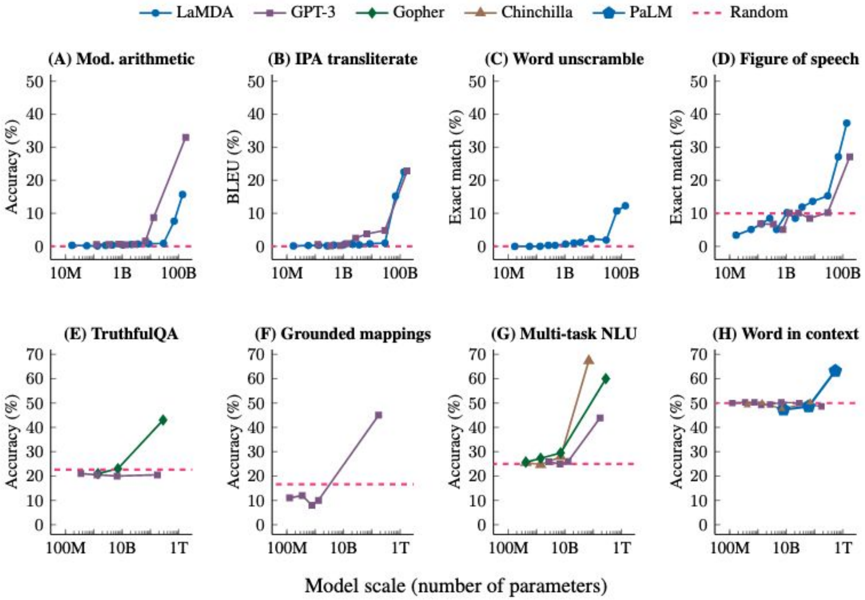
M. de Lhoneux

D. Elliott

Warning: The final part of the section contains dataset samples that are racist in nature.

NLP in the Era of Scale



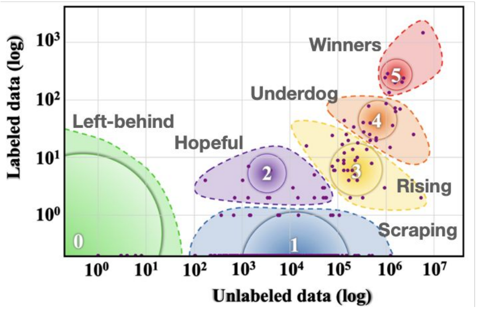


Treviso et al. 2023. Efficient Methods for Natural Language Processing: A Survey. TACL
Wei et al. 2022. Emergent Abilities of Large Language Models. TMLR

NLP for **All** Written Languages

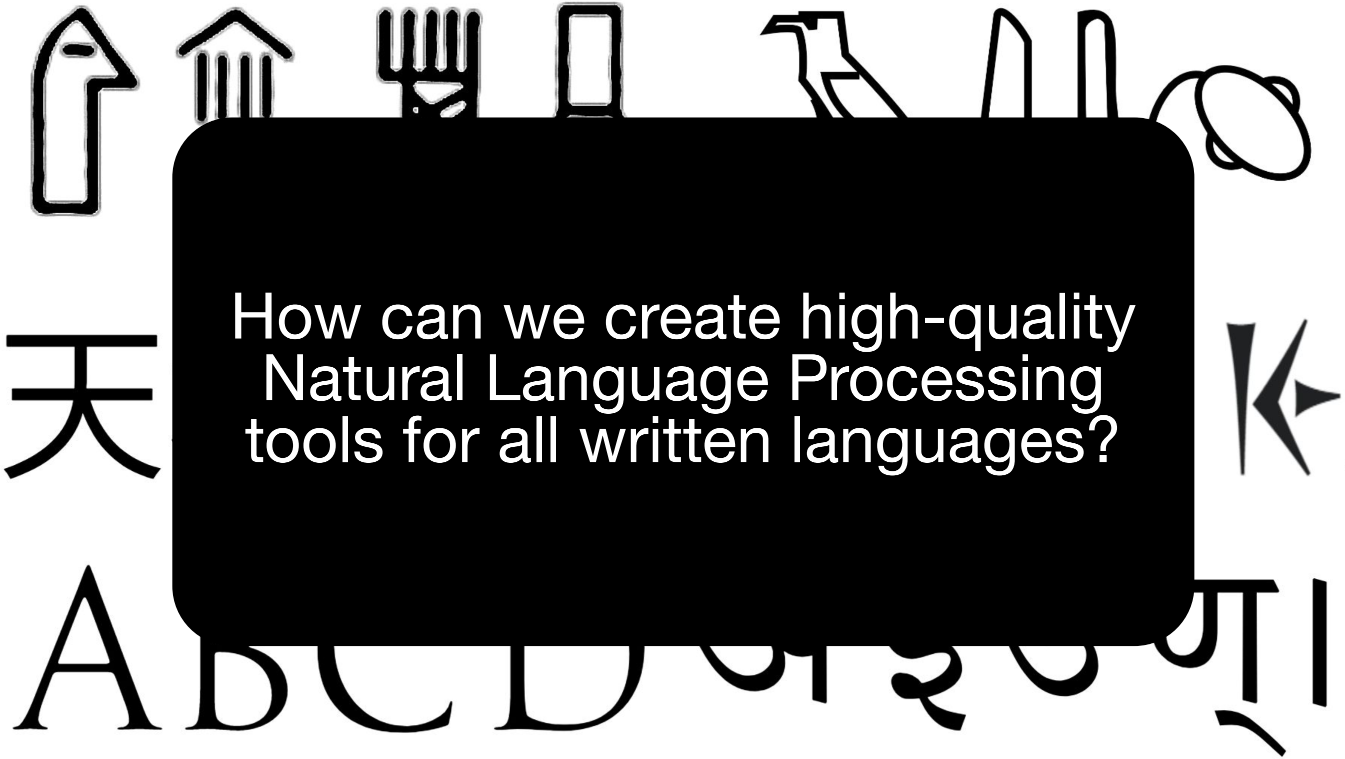
- There are 3,000 written languages
  - 400 with >1M speakers
- NLP usually covers 100 languages
  - Technological exclusion for billions
- We need systems for all languages, not just those that are high-resource

(Data and groupings from Joshi et al. ACL 2020)



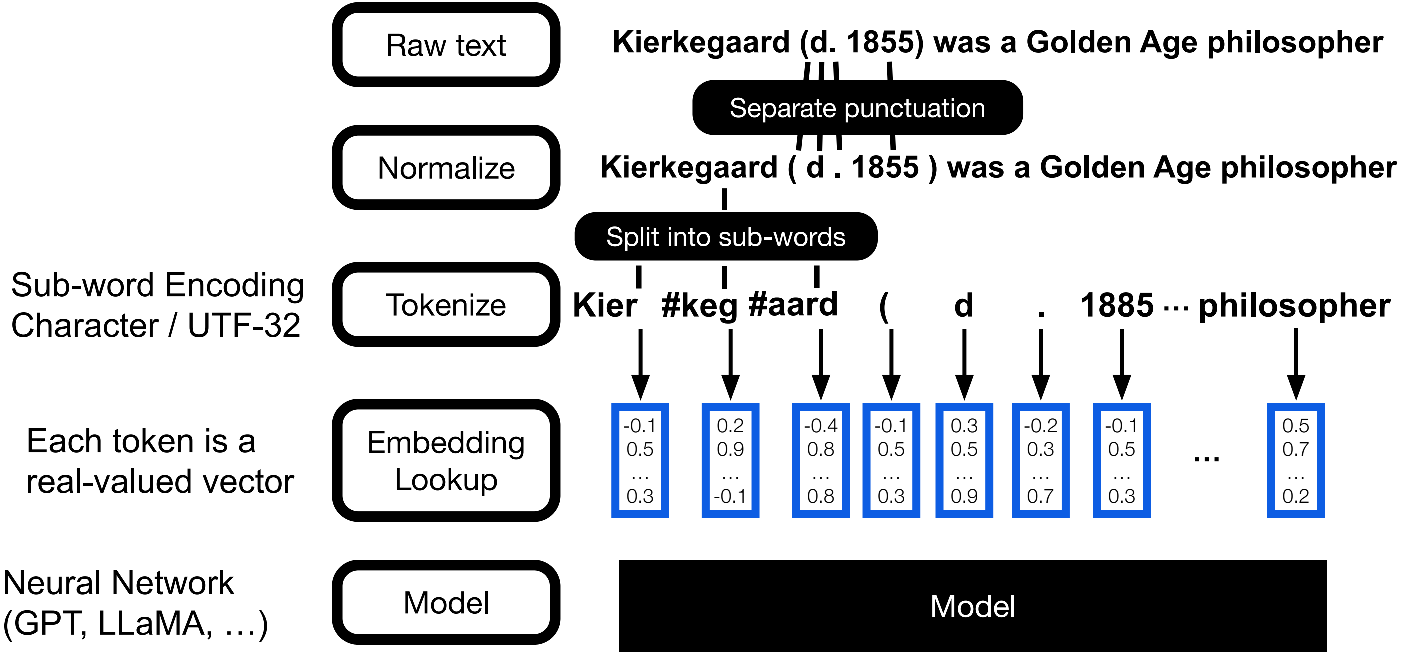
Class	Speakers
0	1.2B
1	30M
2	5.7M
3	1.8B
4	2.2B
5	2.5B

van Esch et al. Writing System and Speaker Metadata for 2,800+ Language Varieties. LREC 2022.
Joshi et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2020.



How can we create high-quality Natural Language Processing tools for all written languages?

NLP is a pipeline …



Raw text

Kierkegaard (d. 1855) was a Golden Age philosopher

Separate punctuation

Normalize

Kierkegaard (d . 1855) was a Golden Age philosopher

Split into sub-words

Sub-word Encoding
Character / UTF-32

Tokenize

Kier #keg #aard (d . 1885 ··· philosopher

Each token is a
real-valued vector

Embedding
Lookup

-0.1 0.5 … 0.3

0.2 0.9 … -0.1

-0.4 0.8 … 0.8

-0.1 0.5 … 0.3

0.3 0.5 … 0.9

-0.2 0.3 … 0.7

-0.1 0.5 … 0.3

…

0.5 0.7 … 0.2

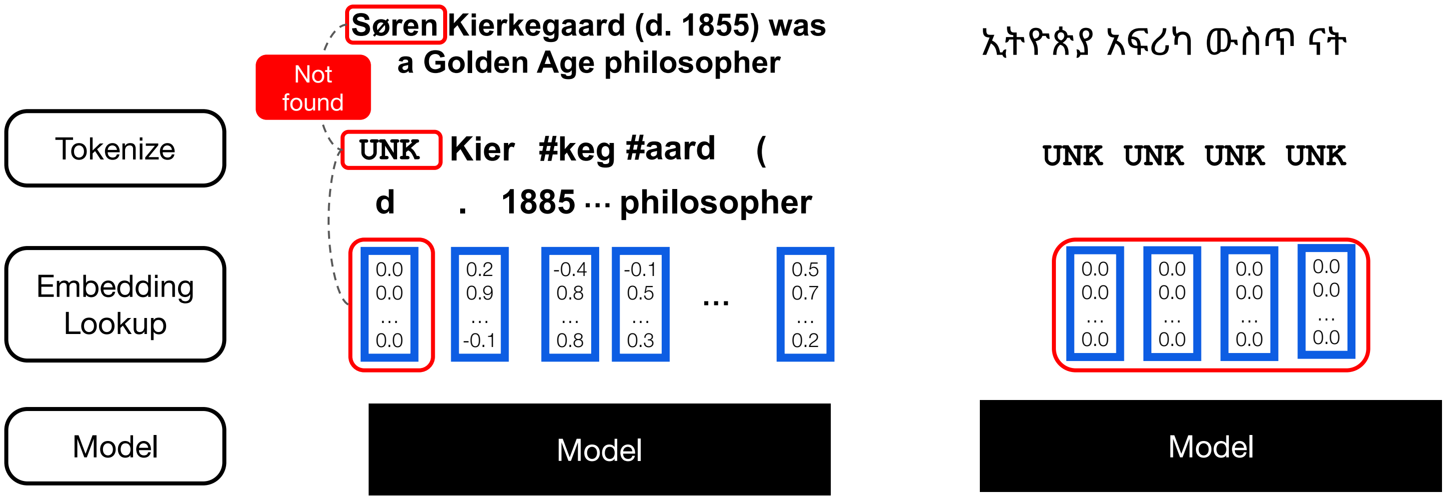
Neural Network
(GPT, LLaMA, …)

Model

Model

Syntactic / Semantic analysis

… that is easily broken



Tokenize

Embedding Lookup

Model

Søren Kierkegaard (d. 1855) was a Golden Age philosopher

Not found

UNK Kier #keg #aard (d . 1885 … philosopher

ኢትዮጵያ አፍሪካ ውስጥ ናት

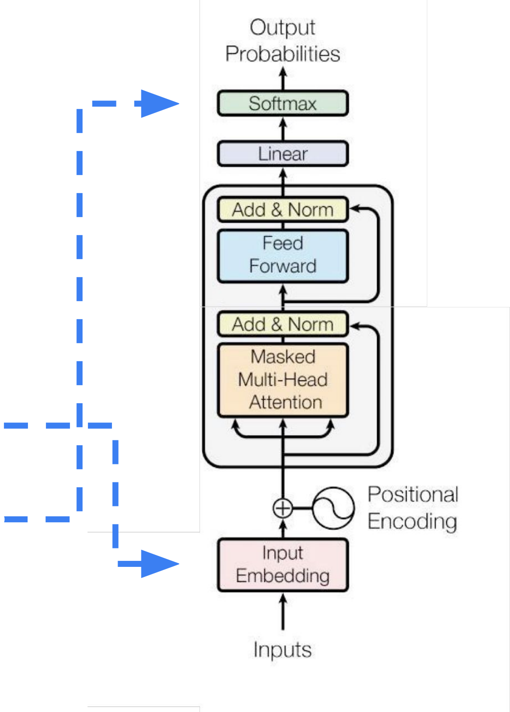
UNK UNK UNK UNK

This issue disproportionately affects low-resource languages

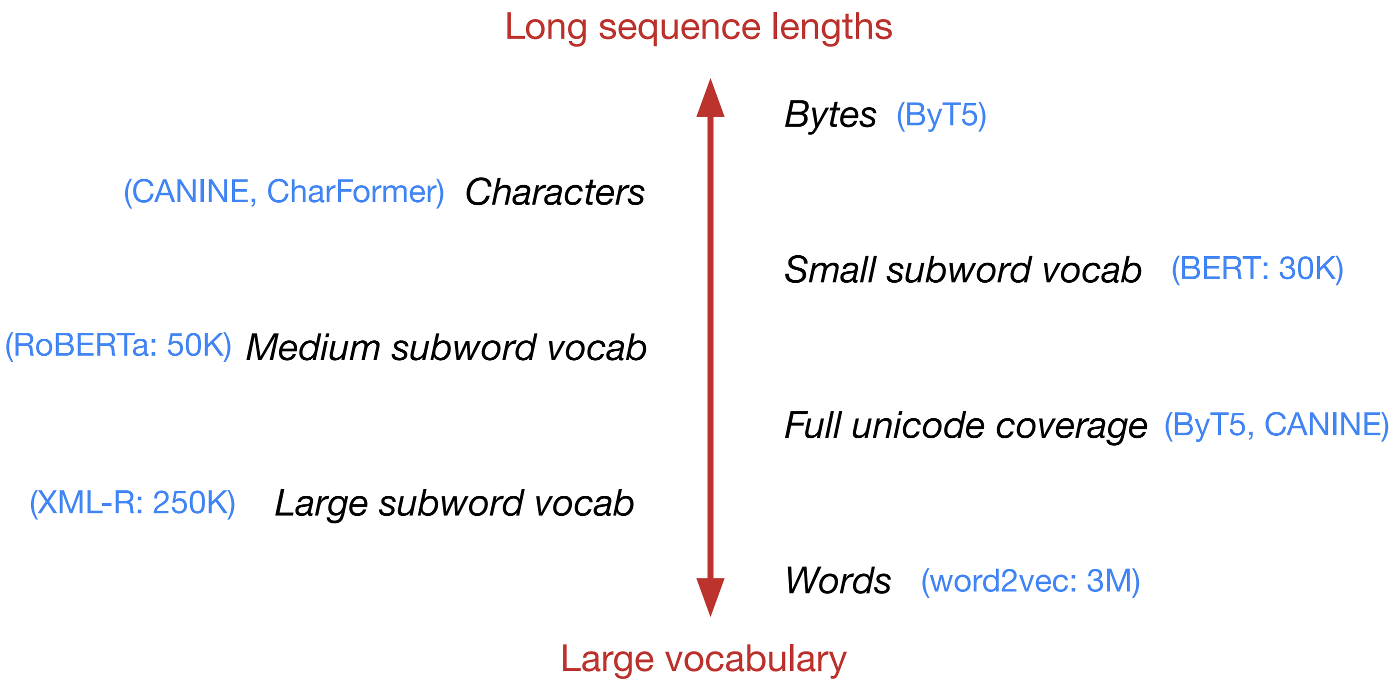
The Vocabulary Bottleneck

- NLP is an **open vocabulary problem** and the ability of a model is determined by its vocabulary:
  1. tokens, characters, sub-words, etc.

- This creates a bottleneck in two places:
  1. *Representational bottleneck* in the Embedding layer
  2. *Computational bottleneck* in the Output layer



Where's the sweet spot?



Long sequence lengths

Bytes (ByT5)

(CANINE, CharFormer) *Characters*

Small subword vocab (BERT: 30K)

(RoBERTa: 50K) *Medium subword vocab*

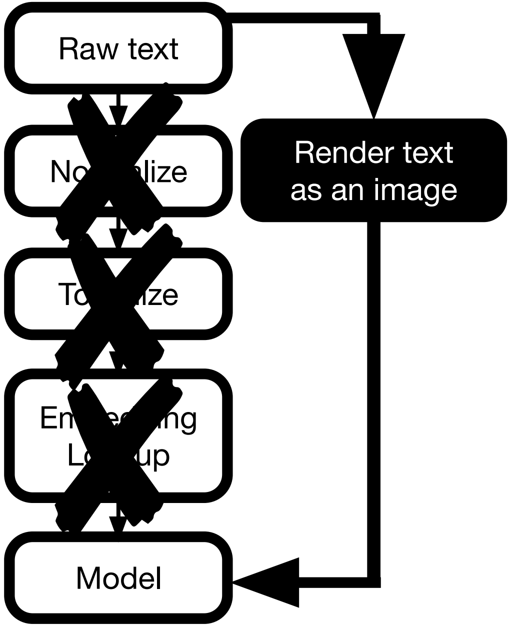
Full unicode coverage (ByT5, CANINE)

(XML-R: 250K) *Large subword vocab*

Words (word2vec: 3M)

Large vocabulary

Main idea: treat language as vision



Raw text

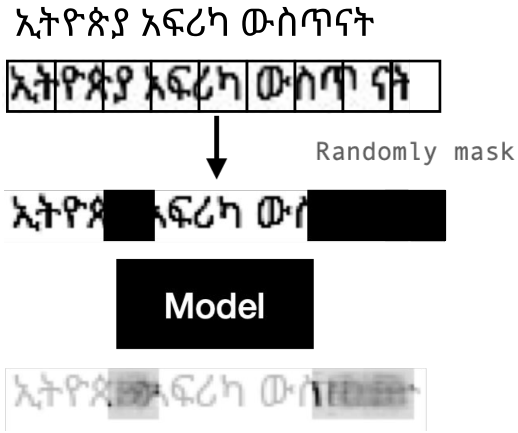
~~Normalize~~

~~Tokenize~~

~~Embedding Lookup~~

Model

Render text as an image



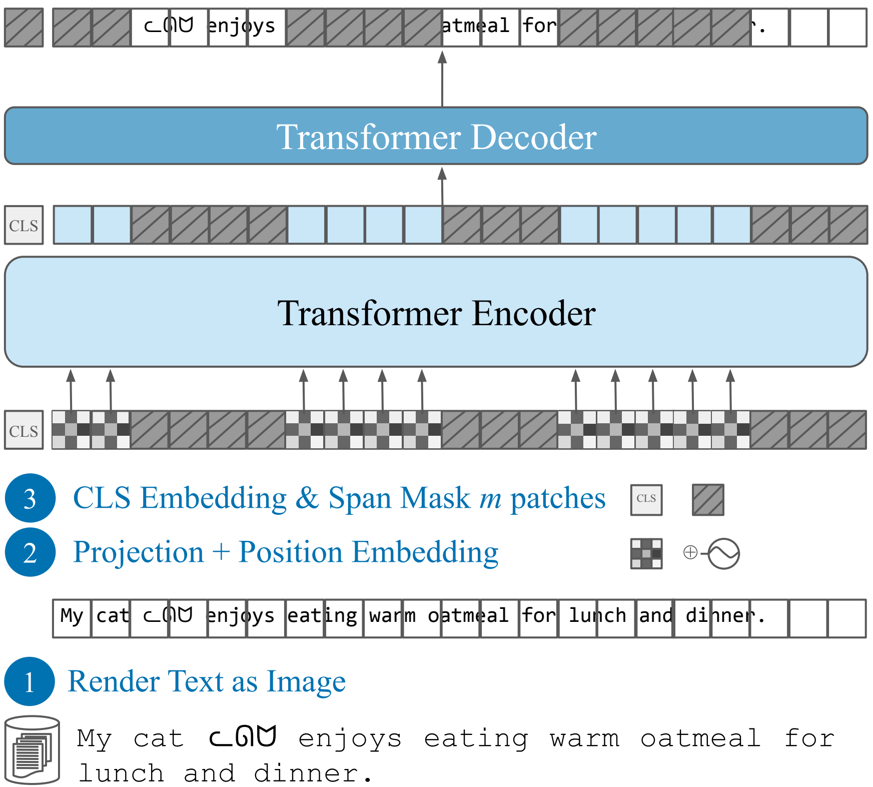
ኢትዮጵያ አፍሪካ ውስጥናት

Randomly mask

Model

The PIXEL Model

$$\frac{1}{M}\sum_{m\in M}(Y_m - \hat{Y}_m | \mathbf{x}_{\setminus m})^2$$



3 CLS Embedding & Span Mask *m* patches CLS ▨

2 Projection + Position Embedding

1 Render Text as Image

My cat ᓚᘏᗢ enjoys eating warm oatmeal for lunch and dinner.

8 Layers

12 Layers

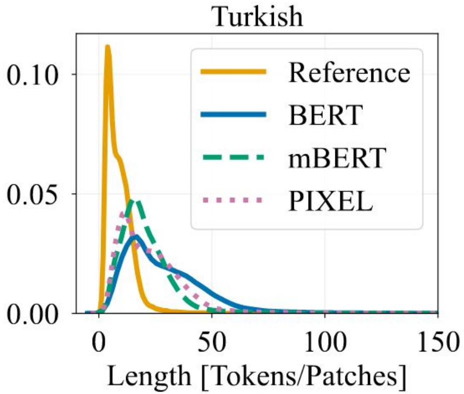
16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

Rendered Text is Compact

- PIXEL encoding produces sequence lengths that are at least as long as as BERT.
  - Universal Dependencies datasets with human reference segmentations
  - No length penalty for any language, unlike some LLMs (Ahia et al. 2023)

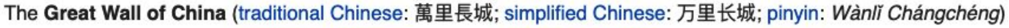


Proportion of text that is encoded as k subwords / patches.

Pretraining

- **English Dataset**: English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length**: 529 patches (`16x8464` pixels)
- **Compute**: 8 x 40GB A100 GPUs for 8 days
- **Parameters**: 86M encoder + 26M decoder

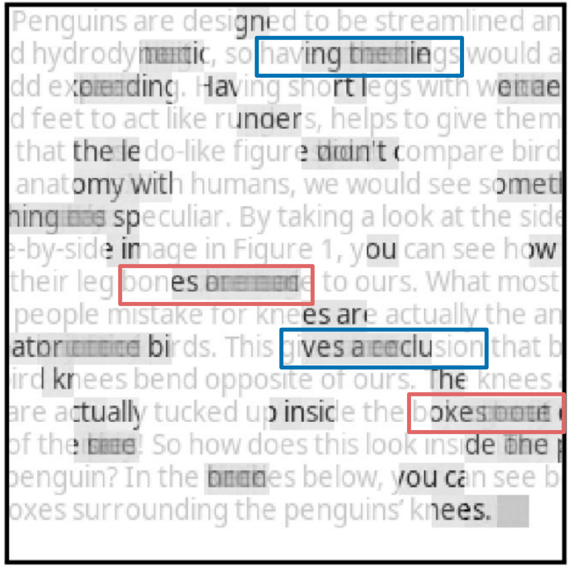
There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)



The **Great Wall of China** (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: *Wànlǐ Chángchéng*)

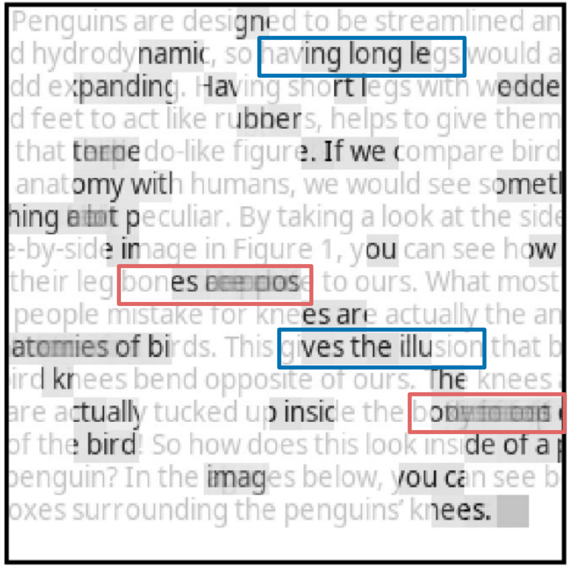
`Pretrained model: https://huggingface.co/Team-PIXEL/pixel-base`

A new type of generative model



Penguins are designed to be streamlined and hydrodynettic, so having theriings would add expleeding. Having short legs with we[illegible] d feet to act like runders, helps to give them that the le do-like figure didn't compare bird anatomy with humans, we would see someth hing [illegible] speculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones [illegible] to ours. What most people mistake for knees are actually the an ator [illegible] birds. This gives a eeclusion that bird knees bend opposite of ours. The knees are actually tucked up inside the bokes [illegible] of the [illegible]! So how does this look inside the p enguin? In the [illegible]es below, you can see b oxes surrounding the penguins' knees.

Try for yourself



Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with wedde d feet to act like rubbers, helps to give them that torpe do-like figure. If we compare bird anatomy with humans, we would see someth hing a lot peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones [illegible]pos to ours. What most people mistake for knees are actually the an atomies of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the body [illegible] of the bird! So how does this look inside of a p enguin? In the images below, you can see b oxes surrounding the penguins' knees.

100K steps → **500K steps** → **1M steps**

Downstream Tasks

- **Datasets**: Universal Dependencies, MasakhaNER, GLUE, Zeroé

- **Models:**

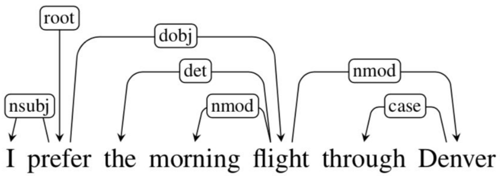
	Parameters	Pretraining Data
$\text{PIXEL}_{\text{BASE}}$	86M	English Wikipedia + Bookcorpus
$\text{BERT}_{\text{BASE}}$	110M	—
CANINE-C	127M	104-languages from Wikipedia

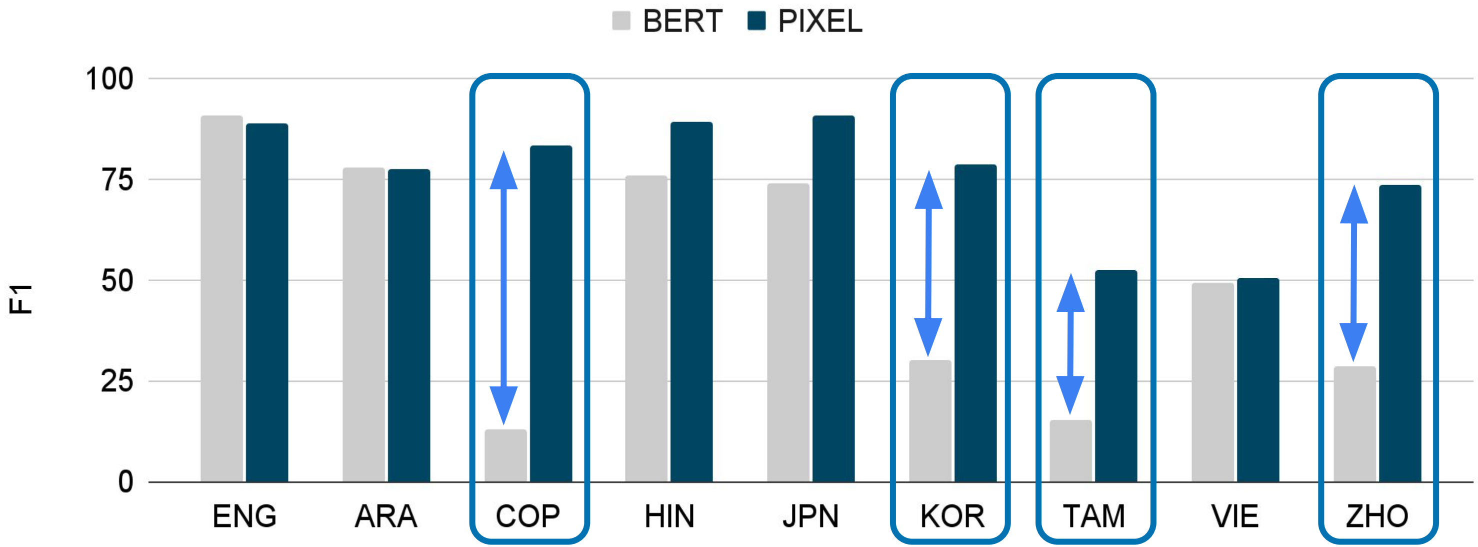
Similar pretraining setup

Tries to solve the same problem using UTF-32

`Code: https://github.com/xplip/pixel`

Dependency Parsing Results



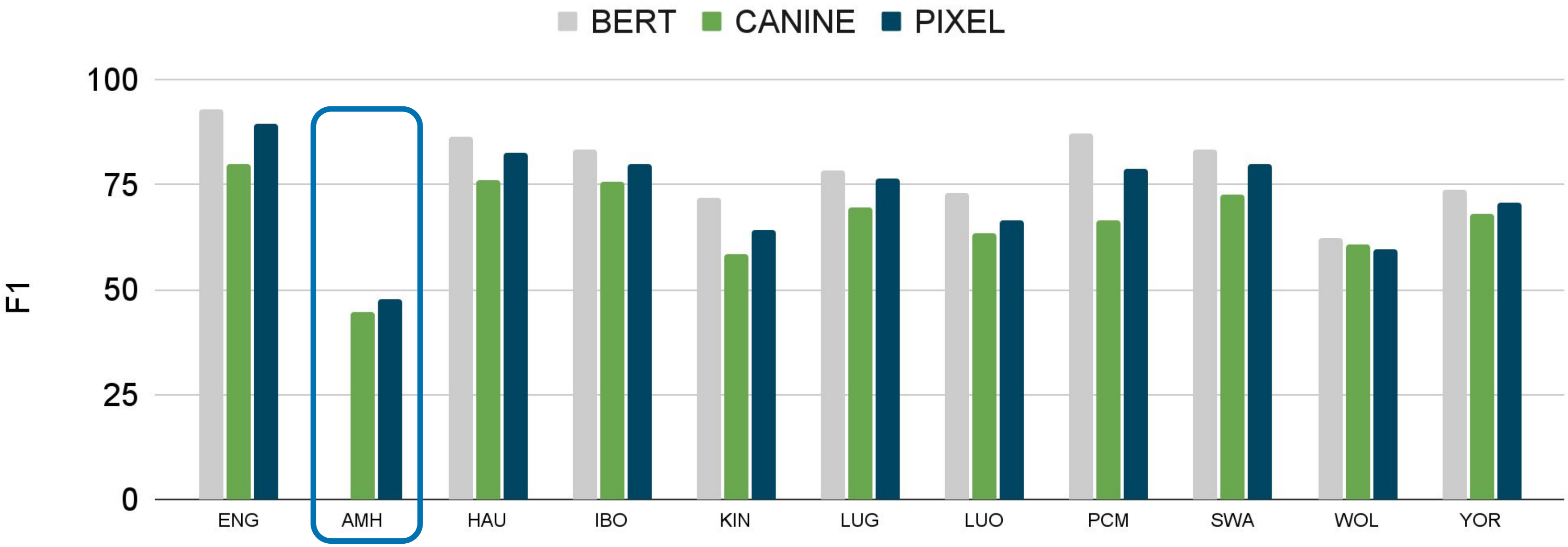


BERT UNK	**0%**	**1%**	**94%**	**33%**	**46%**	**85%**	**82%**	**5%**	**73%**

PIXEL vastly outperforms BERT on unseen scripts

Named Entity Recognition

Emir of Kano turban Zhang wey don spend 18 years for Nigeria



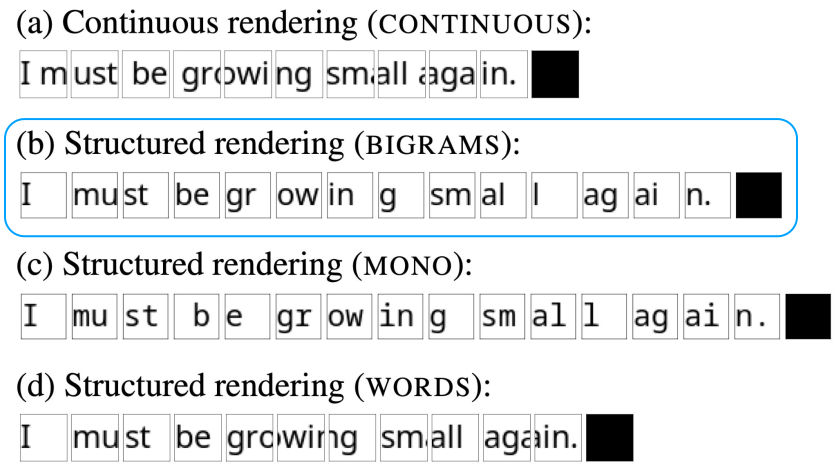
PIXEL outperforms BERT on the non-Latin script

PIXEL outperforms the multilingually pretrained CANINE-C

Text Rendering Matters

- The original text renderer produces many nearly-identical patches
  - This is representation- and compute-wasteful

the
the
the
the
the
the
he
the
he


(a) Continuous rendering (CONTINUOUS):

I must be growing small again.

(b) Structured rendering (BIGRAMS):

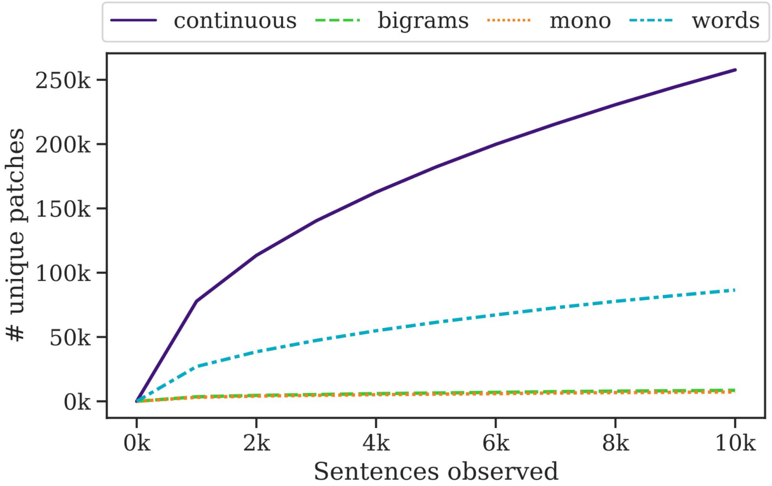
I mu st be gr ow in g sm al l ag ai n.

(c) Structured rendering (MONO):

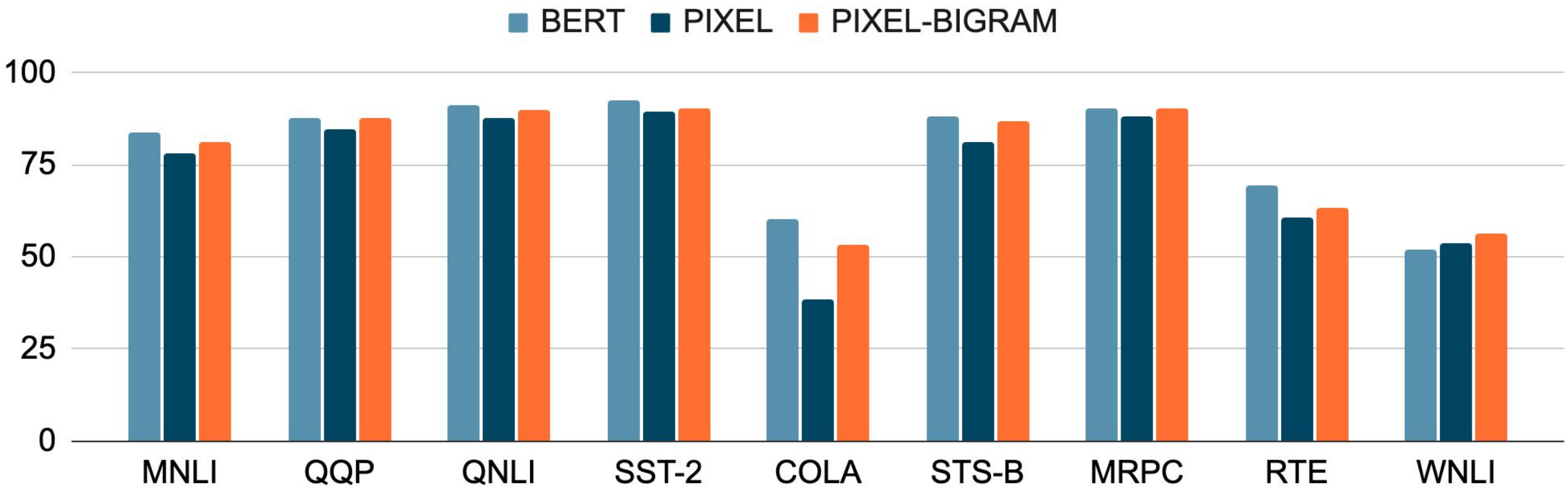
I mu st b e gr ow in g sm al l ag ai n.

(d) Structured rendering (WORDS):

I must be growing small again.



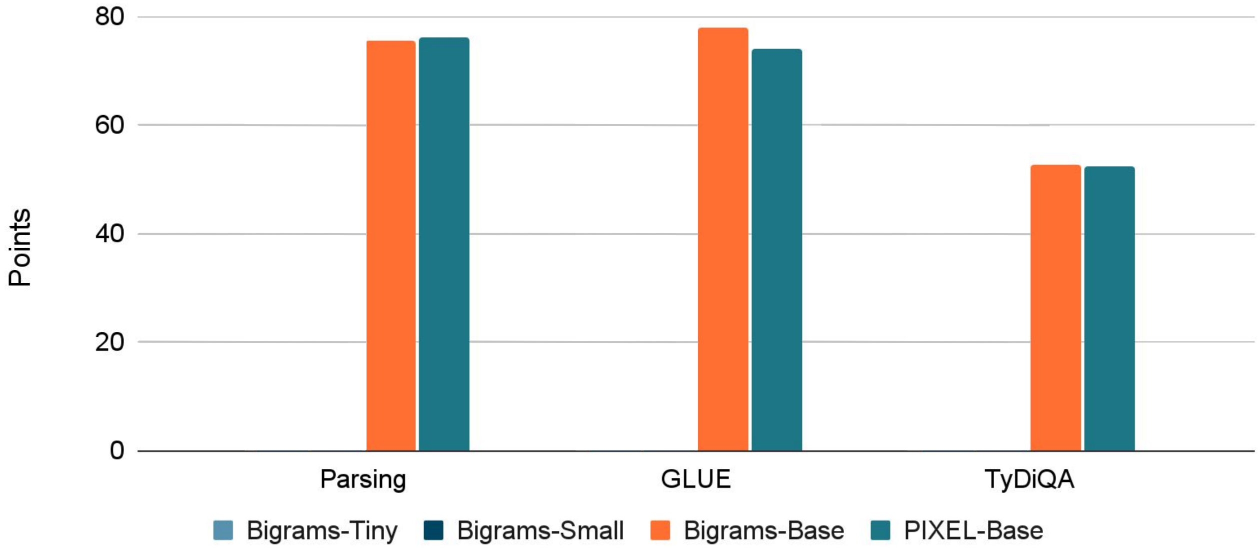
Sentence-level Tasks: GLUE



Bigram text rendering produces better models

Scaling Down ↓

- Better text rendering can create effective models at smaller scales



5M 22M 86M 86M

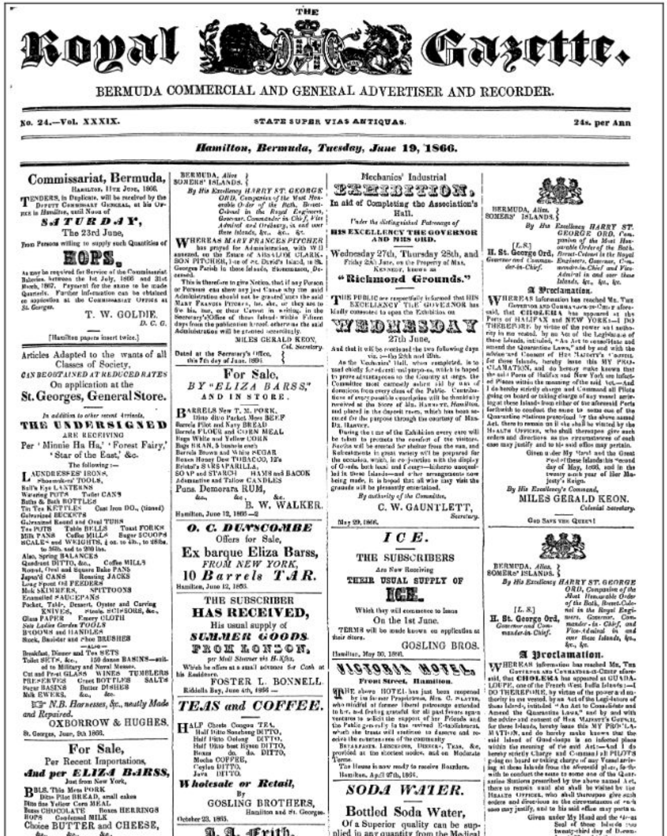
Application: Historical Document Processing

- Worldwide efforts to digitize historic documents (Groesen 2015)
- Typical pipeline for enabling access is:
    a. Scan documents into high-quality digital formats
    b. Perform OCR on those documents (one-off process)
    c. Search through documents using OCR annotations

What if we could do this without OCR?

Caribbean Newspapers, 1718–1876

- Collaboration with researchers that are interested in tracking newspapers notices about escaped slaves
  - What was the given name?
  - What reward was offered?
  - Who was the contact person?
- Dataset of 1.65M scanned pages



THE

Royal Gazette.

BERMUDA COMMERCIAL AND GENERAL ADVERTISER AND RECORDER.

No. 24.—Vol. XXXIX. STATE SUPER VIAS ANTIQUAS. 24s. per Ann

Hamilton, Bermuda, Tuesday, June 19, 1866.

Commissariat, Bermuda,
Hamilton, 11th June, 1866.

TENDERS, in Duplicate, will be received by the Deputy Commissary General, at his Office in Hamilton, until Noon of

SATURDAY,
The 23rd June,

From Persons willing to supply such Quantities of

HOPS,

As may be required for Service of the Commissariat Bakeries, between the 1st July, 1866 and 31st March, 1867. Payment for the same to be made Quarterly. Further information can be obtained on application at the Commissariat Office at St. Georges.

T. W. GOLDIE.
D. C. G.

[Hamilton papers insert twice.]

Articles Adapted to the wants of all Classes of Society,
CAN BE OBTAINED AT REDUCED RATES
On application at the
St. Georges, General Store.

In addition to other recent Arrivals,
THE UNDERSIGNED
ARE RECEIVING
Per 'Minnie Ha Ha,' 'Forest Fairy,' 'Star of the East,' &c.
The following :—

LAUNDRESSES' IRONS,
Shoemakers' TOOLS,
Bull's Eye LANTERNS
Watering POTS Toilet CANS
Baths & Bath BOTTLES
Tin Tea KETTLES Cast Iron DO., (tinned)
Galvanized BUCKETS
Galvanized Round and Oval TUBS
Tea POTS Table BELLS Toast FORKS
Milk PANS Coffee MILLS Sugar SCOOPS
SCALES and WEIGHTS, ¼ oz. to 4lb., to 28lbs. to 56lb. and to 200 lbs.
Also, Spring BALANCES
Quadrant DITTO, &c., Coffee MILLS
Round, Oval and Square Bake PANS
Japan'd CANS Roasting JACKS
Long Spout Oil FEEDERS
Milk SKIMMERS, SPITTOONS
Enamelled SAUCEPANS
Pocket, Table, Dessert, Oyster and Carving KNIVES, Stools, SCISSORS, &c.,
Glass PAPER Emery CLOTH
Sets Ladies Garden TOOLS
BROOMS and HANDLES
Stock, Banister and Shoe BRUSHES
—ALSO—
Breakfast, Dinner and Tea SETS
Toilet SETS, &c., 150 dozen BASINS—suited to Military and Naval Messes.
Cut and Prest GLASS WINES TUMBLERS
PRESERVES Cruet BOTTLES SALTS
Sugar BASINS Butter DISHES
Milk EWERS, &c., &c.

☞ N.B. Harnesses, &c., neatly Made and Repaired.

OXBORROW & HUGHES.
St. Georges, June, 9th 1866.

For Sale,
Per Recent Importations,
And per ELIZA BARSS,
Just from New York,

BBLS. Thin Mess PORK
Ditto Pilot BREAD, small cakes
Ditto fine Yellow Corn MEAL
Boxes CHOCOLATE Boxes HERRINGS
HOPS Condensed MILK
Choice BUTTER and CHEESE,
&c. &c. &c.

BERMUDA, Alias SOMERS' ISLANDS. }

By His Excellency HARRY ST. GEORGE ORD, Companion of the Most Honorable Order of the Bath, Brevet-Colonel in the Royal Engineers, Governor, Commander in Chief, Vice Admiral and Ordinary, in and over these Islands, &c., &c., &c.

WHEREAS MARY FRANCES PITCHER has prayed for Administration, with Will annexed, on the Estate of ABSALOM CLARKSON PITCHER, late of St. David's Island, in St. Georges Parish in these Islands, Stonemason, Deceased.

This is therefore to give Notice, that if any Person or Persons can shew any just Cause why the said Administration should not be granted unto the said Mary Frances Pitcher, he, she, or they are to file his, her, or their Caveat in writing, in the Secretary's Office of these Islands within Fifteen days from the publication hereof, otherwise the said Administration will be granted accordingly.

MILES GERALD KEON,
Col. Secretary.

Dated at the Secretary's Office, this 7th day of June, 1866. }

For Sale,
BY "ELIZA BARSS,"
AND IN STORE.

BARRELS New T. M. PORK,
Ditto ditto Packet Mess BEEF
Barrels Pilot and Navy BREAD
Barrels FLOUR and CORN MEAL
Bags White and Yellow CORN
Bags BRAN, 5 bushels each
Barrels Brown and White SUGAR
Boxes Honey Dew TOBACCO, 12's
Bristol's SARSAPARILLA,
SOAP and STARCH HAMS and BACON
Adamantine and Tallow CANDLES
Puns. Demerara RUM,
&c., &c., &c.

B. W. WALKER.
Hamilton, June 12, 1866—2

O. C. DUNSCOMBE
Offers for Sale,
Ex barque Eliza Barss,
FROM NEW YORK,
10 Barrels TAR.
Hamilton, June 12, 1866.

THE SUBSCRIBER
HAS RECEIVED,
His usual supply of
SUMMER GOODS
FROM LONDON,
per Mail Steamer via Halifax,

Which he offers at a small advance for Cash at his Residence.

FOSTER L. BONNELL.
Riddells Bay, June 4th, 1866 —

TEAS and COFFEE.

HALF Chests Congou TEA,
Half Ditto Souchong DITTO,
Half Ditto Oolong DITTO,
Half Ditto best Hyson DITTO,
Boxes do. do. DITTO,
Mocha COFFEE,
Ceylon DITTO,
Java DITTO.

Wholesale or Retail,
By
GOSLING BROTHERS,
Hamilton and St. Georges.
October 23, 1865.

B. A. Frith.

Mechanics' Industrial
EXHIBITION,
In aid of Completing the Association's Hall.
Under the distinguished Patronage of
HIS EXCELLENCY THE GOVERNOR AND MRS ORD.

Wednesday 27th, Thursday 28th, and Friday 29th June, on the Property of Mrs. Kennedy, known as
"Richmond Grounds."

THE PUBLIC are respectfully informed that HIS EXCELLENCY THE GOVERNOR has kindly consented to open the Exhibition on

WEDNESDAY
27th June,

And that it will be continued the two following days viz. :—the 28th and 29th.

As the Mechanics' Hall, when completed, is to used chiefly for educational purposes, which is hoped to prove advantageous to the Country at large, the Committee most earnestly solicit aid by way of donations from every class of the Public. Contributions of every possible description will be thankfully received at the Store of Mr. Harnett, Hamilton, and placed in the deposit room, which has been secured for the purpose through the courtesy of Mrs. Dr. Harvey.

During the time of the Exhibition every care will be taken to promote the comfort of the visitors. Booths will be erected for shelter from the sun, and Refreshments in great variety will be prepared for the occasion, which, in conjunction with the display of Goods, both local and foreign—hitherto unequalled in these Islands—and other arrangements now being made, it is hoped that all who may visit the grounds will be pleasantly entertained.

By authority of the Committee,
C. W. GAUNTLETT,
Secretary.
May 29, 1866.

ICE.
THE SUBSCRIBERS
Are Now Receiving
THEIR USUAL SUPPLY OF
ICE.
Which they will commence to Issue
On the 1st June.

TERMS will be made known on application at their Store.

GOSLING BROS.
Hamilton, May 30, 1866.

VICTORIA HOTEL,
Front Street, Hamilton.

THE above HOTEL has just been reopened by its former Proprietress, Mrs. C. Slater, who mindful of former liberal patronage extended to her, and feeling grateful for all past favors again ventures to solicit the support of her Friends and the Public generally in the revived Establishment, which she trusts will continue to deserve and receive the countenance of the community.

Breakfasts, Luncheons, Dinners, Teas, &c, provided at the shortest notice, and on Moderate Terms.

The House is now ready to receive Boarders.

Hamilton, April 27th, 1866.

SODA WATER.

Bottled Soda Water,

Of a Superior quality can be supplied in any quantity from the Medical

BERMUDA, Alias, SOMERS' ISLANDS. }

By His Excellency HARRY ST. GEORGE ORD, Companion of the Most Honorable Order of the Bath, Brevet-Colonel in the Royal Engineers, Governor, Commander-in-Chief and Vice-Admiral in and over these Islands, &c., &c., &c.

[L.S.]
H. St. George Ord,
Governor and Commander-in-Chief.

A Proclamation.

WHEREAS information has reached Me, The Governor and Commander-in-Chief aforesaid, that CHOLERA has appeared at the Ports of HALIFAX and NEW YORK:—I DO THEREFORE, by virtue of the power and authority in me vested, by an Act of the Legislature of these Islands, intituled, "An Act to consolidate and amend the Quarantine Laws," and by and with the advice and Consent of Her Majesty's Council for these Islands, hereby issue this MY PROCLAMATION, and do hereby make known that the said Ports of Halifax and New York are Infected Places within the meaning of the said Act.—And I do hereby strictly charge and Command all Pilots going on board or taking charge of any vessel arriving at these Islands from either of the aforesaid Ports forthwith to conduct the same to some one of the Quarantine Stations prescribed by the above named Act, there to remain until she shall be visited by the Health Officer, who shall thereupon give such orders and directions as the circumstances of each case may justify and to his said office may pertain.

Given under My Hand and the Great Seal of these Islands this Second day of May, 1866, and in the twenty ninth year of Her Majesty's Reign.

By His Excellency's Command,
MILES GERALD KEON.
Colonial Secretary.

God Save the Queen!

BERMUDA, Alias, SOMERS' ISLANDS. }

By His Excellency HARRY ST. GEORGE ORD, Companion of the Most Honorable Order of the Bath, Brevet-Colonel in the Royal Engineers, Governor, Commander-in-Chief, and Vice-Admiral in and over these Islands, &c., &c., &c.

[L. S.]
H. St. George Ord,
Governor and Commander-in-Chief.

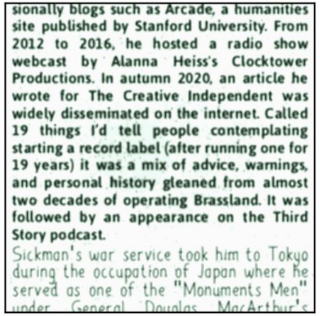
A Proclamation.

WHEREAS information has reached Me, The Governor and Commander-in-Chief aforesaid, that CHOLERA has appeared at GUADALOUPE, one of the French West India Islands:—I DO THEREFORE, by virtue of the power and authority in me vested, by an Act of the Legislature of these Islands, intituled "An Act to Consolidate and Amend the Quarantine Laws," and by and with the advice and consent of Her Majesty's Council for these Islands, hereby issue this MY PROCLAMATION, and do hereby make known that the said Island of Guadaloupe is an infected place within the meaning of the said Act:—And I do hereby strictly Charge and Command all PILOTS going on board or taking charge of any Vessel arriving at these Islands from the aforesaid place, forthwith to conduct the same to some one of the Quarantine Stations prescribed by the above named Act, there to remain until she shall be visited by the Health Officer, who shall thereupon give such orders and directions as the circumstances of each case may justify, and to his said office may pertain.

Given under My Hand and the Great Seal of these Islands this twenty-third day of Decem-

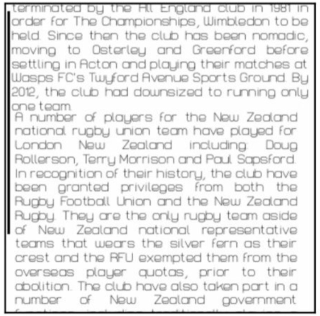
PIXEL for Historical Documents

- Historical document-aware Pretraining
  - Mixture of scanned newspapers and synthetic newspaper-like text generated from Wikipedia and Bookcorpus datasets
  - All input data is scaled to 368x368 and split into 16x16 patches



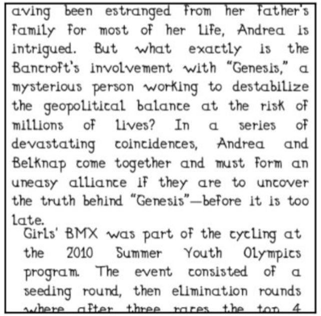
sionally blogs such as Arcade, a humanities site published by Stanford University. From 2012 to 2016, he hosted a radio show webcast by Alanna Heiss's Clocktower Productions. In autumn 2020, an article he wrote for The Creative Independent was widely disseminated on the internet. Called 19 things I'd tell people contemplating starting a record label (after running one for 19 years) it was a mix of advice, warnings, and personal history gleaned from almost two decades of operating Brassland. It was followed by an appearance on the Third Story podcast.

Sickman's war service took him to Tokyo during the occupation of Japan where he served as one of the "Monuments Men" under General Douglas MacArthur's



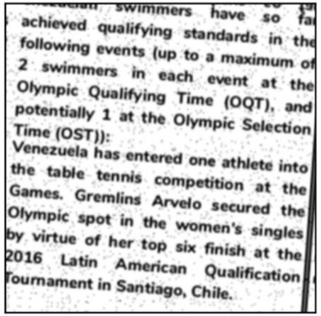
terminated by the All England club in 1981 in order for The Championships, Wimbledon to be held. Since then the club has been nomadic, moving to Osterley and Greenford before settling in Acton and playing their matches at Wasps FC's Twyford Avenue Sports Ground. By 2012, the club had downsized to running only one team.

A number of players for the New Zealand national rugby union team have played for London New Zealand including: Doug Rollerson, Terry Morrison and Paul Sapsford. In recognition of their history, the club have been granted privileges from both the Rugby Football Union and the New Zealand Rugby. They are the only rugby team aside of New Zealand national representative teams that wears the silver fern as their crest and the RFU exempted them from the overseas player quotas, prior to their abolition. The club have also taken part in a number of New Zealand government



aving been estranged from her father's family for most of her life, Andrea is intrigued. But what exactly is the Bancroft's involvement with "Genesis," a mysterious person working to destabilize the geopolitical balance at the risk of millions of lives? In a series of devastating coincidences, Andrea and Belknap come together and must form an uneasy alliance if they are to uncover the truth behind "Genesis"—before it is too late.

Girls' BMX was part of the cycling at the 2010 Summer Youth Olympics program. The event consisted of a seeding round, then elimination rounds where after three races the top 4

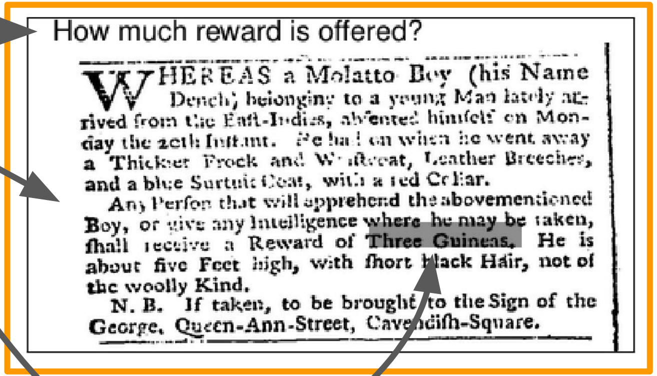


swimmers have so far achieved qualifying standards in the following events (up to a maximum of 2 swimmers in each event at the Olympic Qualifying Time (OQT), and potentially 1 at the Olympic Selection Time (OST)):

Venezuela has entered one athlete into the table tennis competition at the Games. Gremlins Arvelo secured the Olympic spot in the women's singles by virtue of her top six finish at the 2016 Latin American Qualification Tournament in Santiago, Chile.

Visual Question Answering in Newspapers

- Frame this as a Visual Question Answering Task
  - Render the question
  - Render the clipping on a canvas
  - Annotate context of answer
- Train the model to predict the label of the answer



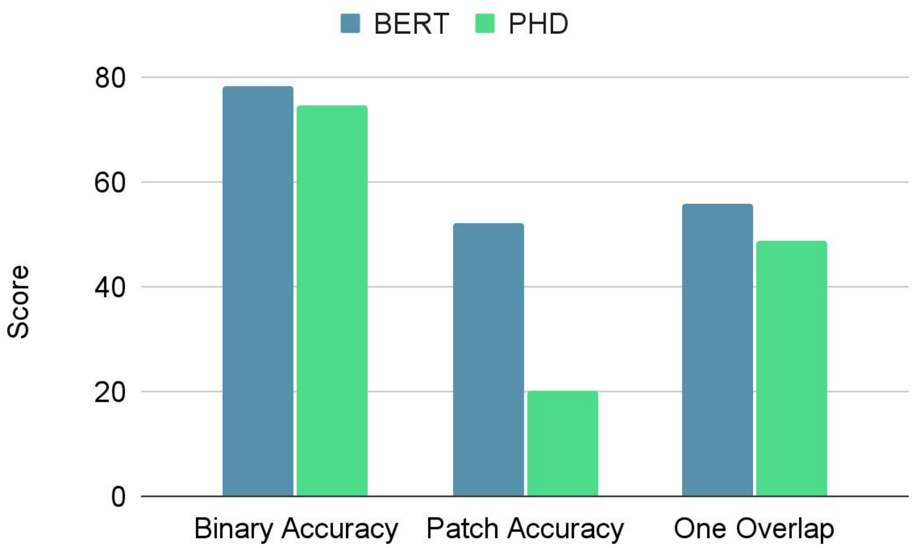
How much reward is offered?

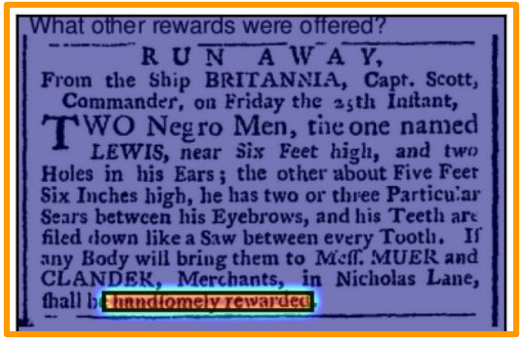
WHEREAS a Molatto Boy (his Name Dench) belonging to a young Man lately arrived from the Eaſt-Indies, abſented himſelf on Monday the 20th Inſtant. He had on when he went away a Thickset Frock and Waiſtcoat, Leather Breeches, and a blue Surtout Coat, with a red Collar.

Any Perſon that will apprehend the abovementioned Boy, or give any Intelligence where he may be taken, ſhall receive a Reward of Three Guineas. He is about five Feet high, with ſhort black Hair, not of the woolly Kind.

N. B. If taken, to be brought to the Sign of the George, Queen-Ann-Street, Cavendiſh-Square.

Results



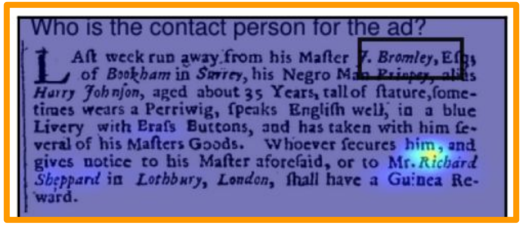


What other rewards were offered?

RUN AWAY,

From the Ship BRITANNIA, Capt. Scott, Commander, on Friday the 25th Inſtant,

TWO Negro Men, the one named LEWIS, near Six Feet high, and two Holes in his Ears; the other about Five Feet Six Inches high, he has two or three Particular Sears between his Eyebrows, and his Teeth are filed down like a Saw between every Tooth. If any Body will bring them to Meſſ. MUER and CLANDER, Merchants, in Nicholas Lane, ſhall be handſomely rewarded.



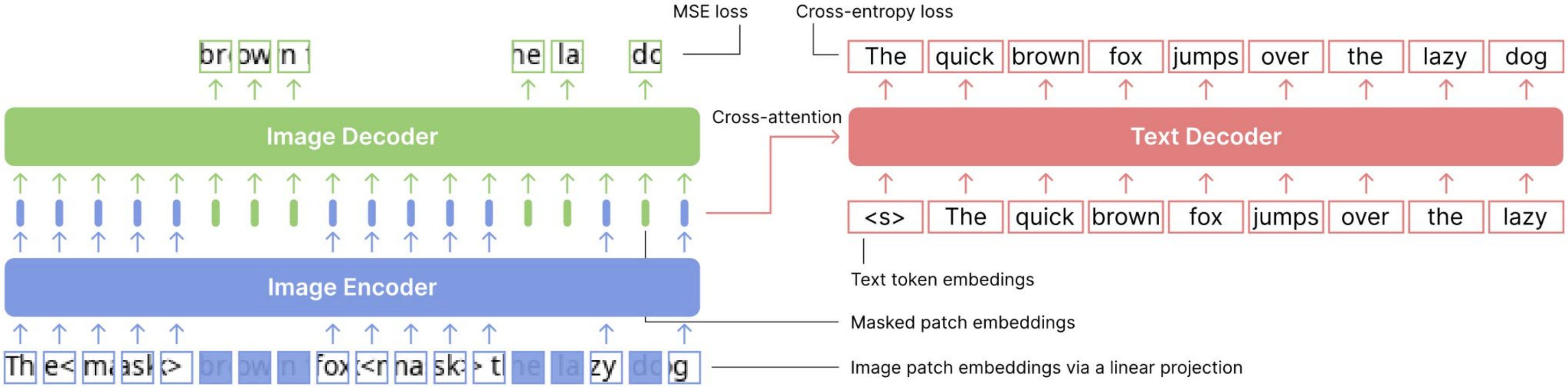
Who is the contact person for the ad?

LAſt week run away from his Maſter J. Bromley, Eſq; of Bookham in Surrey, his Negro Man Prince, alias Harry Johnſon, aged about 35 Years, tall of ſtature, ſometimes wears a Perriwig, ſpeaks Engliſh well, in a blue Livery with Braſs Buttons, and has taken with him ſeveral of his Maſters Goods. Whoever ſecures him, and gives notice to his Maſter aforeſaid, or to Mr. Richard Sheppard in Lothbury, London, ſhall have a Guinea Reward.

Surprisingly good performance compared to a model trained on manually transcribed text

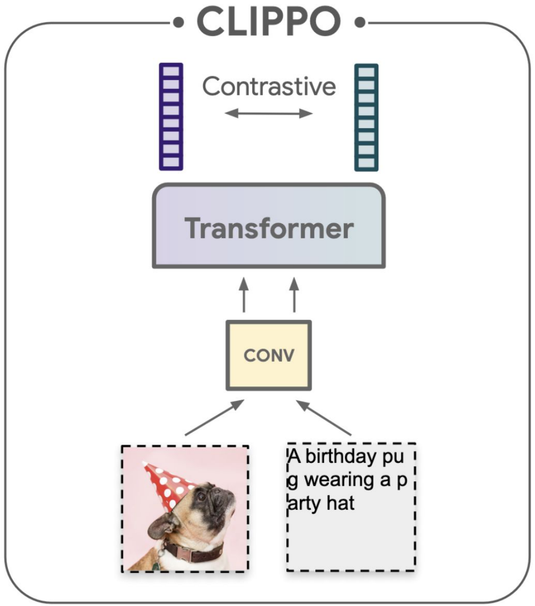
Patch and Text Prediction

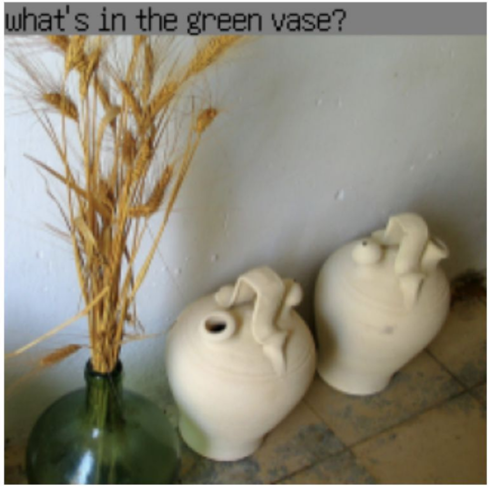
- Combine patch and token prediction


MSE loss
Cross-entropy loss
br ow n t
he la
do
The quick brown fox jumps over the lazy dog
Image Decoder
Cross-attention
Text Decoder
<s> The quick brown fox jumps over the lazy
Text token embedings
Image Encoder
Masked patch embeddings
Th e< ma ask <> br ow n fox <r na sk> tl he la zy do g
Image patch embeddings via a linear projection

Gao et al. 2024. Improving Language Understanding from Screenshots

Joint Multimodal Reasoning





Tschannen et al. CVPR 2023. CLIPPO: Image-and-Language Understanding from Pixels Only.

Open Questions

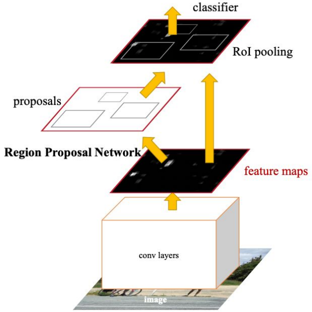
- Interpretability:
    - Does this work based on orthographic similarity or is it learning grammatical representations of text from pixels?
- Multilinguality and scale:
    - How should we train a multilingual PIXEL encoder?
        - Language-based or script-based data selection

Wrap-up

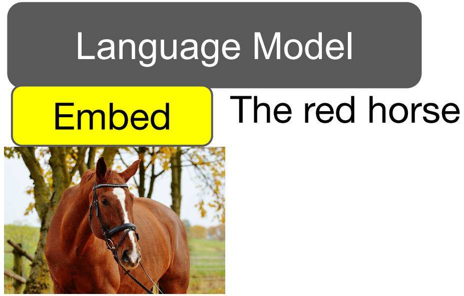
1. Datasets

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.

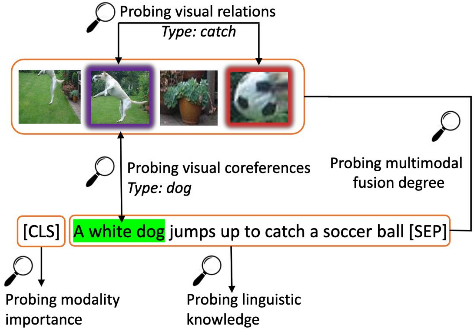
2. Representation



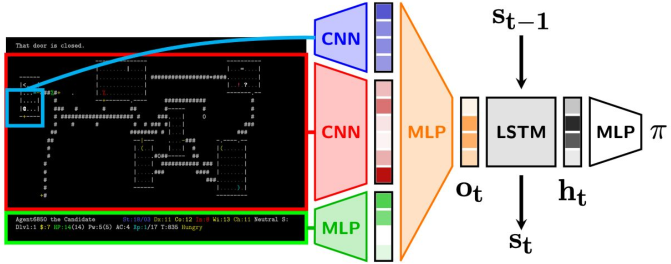
3. Modelling



4. Understanding


Probing visual relations
Type: catch
Probing visual coreferences
Type: dog
Probing multimodal fusion degree
[CLS]
A white dog jumps up to catch a soccer ball [SEP]
Probing modality importance
Probing linguistic knowledge

5. New Directions



```
That door is closed.
```

```
Agent6850 the Candidate        St:18/03 Dx:11 Co:12 In:8 Wi:13 Ch:11 Neutral S:
Dlvl:1 $:7 HP:14(14) Pw:5(5) AC:4 Xp:1/17 T:835 Hungry
```